This article has been accepted for publication in IEEE Transactions on Information Theory. DOI: 10.1109/TIT.2023.3258182. Available at: http://ieeexplore.ieee.org

Regularized Estimation of Information via Canonical Correlation Analysis on a Finite-Dimensional Feature Space

Ferran de Cabrera and Jaume Riba, *Senior Member, IEEE*

***Abstract*—This paper aims to estimate the information between two random phenomena by using consolidated second-order statistics tools. The squared-loss mutual information, a surrogate of the Shannon mutual information, is chosen due to its property of being expressed as a second-order moment. We first review the rationale for i.i.d. discrete sources, which involves mapping the data onto the simplex space, and we highlight the links with other well-known related concepts in the literature based on local approximations of information-theoretic measures. Then, the problem is translated to analog sources by mapping the data onto the characteristic space, focusing on the adaptability between the discrete and the analog case and its limitations. The proposed approach gains interpretability and scalability for its use on large data sets, providing a unified rationale for the free regularization parameters. Moreover, the structure of the proposed mapping allows resorting to Szegö's theorem to reduce the complexity for high dimensional mappings, exhibiting a strong duality with spectral analysis. The performance of the developed estimators is analyzed using Gaussian mixtures.**

***Index Terms*—Data analytics, Canonical correlation analysis, Squared-loss mutual information, Characteristic function, Information-theoretic learning.**

I. Introduction

ENTROPY and mutual information, introduced by Shannon in 1948, are well-known concepts with clear operational significance in the field of information theory and communications that establish fundamental limits in data compression and data transmission [1]. More generally, Kullback-Leibler (KL) divergence (also called relative entropy) is a dissimilarity measure between distributions that includes the notion of mutual information. In the last decades, some researchers have used entropy, mutual information, and divergence in a wide class of areas beyond communications, such as data science, machine learning, neuroscience, economics, biology, language, and other experimental sciences. These quantifiers of information have proven their utility as tools for measuring randomness, dependence, and similarity of random phenomena [2], [3], substituting or working together with the conventional statistical tools of variance and covariance. As a prominent example, the field of information-theoretic learning [4] cuts across signal processing and machine learning by reviewing the learning process under the umbrella of information theory. This new perspective of knowledge discovering provides the guidelines for the design of nonparametric universal tools for data analytics [5]. In particular, the wish for interpretability has become a particularly challenging aspect in practical applications of machine learning systems due to their lack of ability to explain their actions to humans [6]. Although these applications exhibit impressive capabilities, the development of tools for measuring information can provide ways to diagnose in the case of failures.

By delving into fundamental concepts of information theory and statistical signal processing, this paper aims at developing insightful tools for measuring meaningful indicators of the amount of information contained in raw data with the objective of leveraging classical and consolidated statistical signal processing techniques based on second-order statistics. The main contributions of this work are the following:

1) To review the suitability of the squared-loss mutual information surrogate for both discrete and analog sources, focusing on its inherent property of being expressed as a second-order statistic, and contextualizing in terms of similar works. We also review its relationship with mutual information in the small dependence regime.
2) To link the problem of estimating information with the classical problem of Canonical Correlation Analysis (CCA), a widely used tool in many fields of statistical signal processing. By translating the problem of estimating the squared-loss mutual information to the problem of estimating covariance and correlation matrices, the consistency of the estimator is ensured, at least for discrete sources.
3) The proposal of an explicit mapping from analog sources to complex steering vectors based on the characteristic function. This specific transformation leads to a computationally efficient alternative to the Kernel CCA (KCCA) and its mechanism for regularization.
4) To provide interpretability to the problem of estimating information and its regularization, thus endowing the methodology with insight on the selection of its free parameters. The proposed structure for estimation allows for establishing the link between the data size and the degree of regularization without the need to restore to a

This work has been supported by the Spanish Ministry of Science and Innovation through project RODIN (PID2019-105717RB-C22/MCIN/AEI/10.13039/501100011033), by the grant 2021 SGR 01033 (AGAUR, Generalitat de Catalunya), and fellowship FI 2019 by the Secretary for University and Research of the Generalitat de Catalunya and the European Social Fund.

Ferran de Cabrera and Jaume Riba are with the Signal Theory and Communications Department, Universitat Politècnica de Catalunya (UPC), Barcelona 08034, Spain (e-mail: jaume.riba@upc.edu; ferran.de.cabrera@upc.edu).

This paper was presented in part at the 44th International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019.

bias-variance trade-off.

5) The proposal of a reduced complexity approximate estimator resorting to the asymptotic behavior of Toeplitz matrices.

A. Related works and overall organization

While the estimation of information measures such as mutual information is a challenging problem, its literature has a long and rich history. Traditionally, plug-in methods have been widely used for estimating entropy and divergence. These methods are based on first estimating the distributions of the observed data and incorporating them into the functional of interest in a second stage [7]. However, plug-in methods are generally susceptible to estimation errors when dealing with random variables with long tails in their distributions [8]. Estimators based on partitioning the observation space have shown good results by smartly adapting the partitions to the data samples [9], with no requirement of smoothness or tail conditions, but a penalty term to reduce the bias is required. Similarly, the estimators based on the nearest neighbor algorithm provide a different approach on the partitioning criterion, which is particularly useful for high-dimensional data [10]. These estimators perform well for a small choice of k neighbors, but they can be time consuming for large data sets. A more detailed guide of empirical estimators has been recently provided in [3], and for analog sources in particular in [2]. Surrogates of entropy, KL divergence, and mutual information, such as Rényi entropy, Rényi divergence, and χ^2-divergence have also been of particular interest for estimation purposes. The cases of second-order Rényi entropy and divergence cope particularly well with plug-in estimates, which has been shown in multiple applications [4], [11]. Regarding the surrogates of mutual information, the estimation of the Squared-Loss Mutual Information (SMI) has been proposed in [12] by directly estimating the density ratio between the joint and the product of marginal probability density functions, albeit its parameters need to be selected through cross-validation.

Concerning the related works, the presented paper is an extension of the main ideas briefly provided by the authors in [13], where a first approach to estimating the SMI was presented. The idea of local approximations of information measures exposed in this paper is very similar to the linear information coupling approach proposed in [5], [14], which was used there as a tool for developing insights on otherwise intractable problems in the field of communications. The study of the modal decomposition of distributions and the measurement of information through CCA in [15] is parallel to the development provided in this article. However, here we focus on their estimation by proposing a particular feature map in the case of analog sources.

The decomposition of the SMI as the sum of correlation measures between two transformed random variables is provided in [17], called the Principal Inertia Components (PICs). The relationship between the largest PIC and the maximal correlation coefficient is also provided. However, while in [17] the focus is on discrete data and privacy applications, the derivations in this paper concern the estimation of the SMI itself. The estimation of these features is also studied in [16], although focused on the information-theoretic interpretation of deep neural networks.

The proposed method can be cast as a primal form of the KCCA [18], which uses the dual model by means of the kernel trick. Thus, we gain intuition and scalability of the overall data processing by performing a measure of correlation on the feature space rather than in the infinite-dimensional space [19]. This approach is particularly useful if the feature space dimension is smaller than the sample size. Moreover, in [20] it is suggested that the right choice of mapping function leads to a better representation of the data in the feature space, enabling to capture as much information as possible with a reduced dimensional mapping. The proposed statistic based on a Frobenius norm of a coherence matrix is related to the local test proposed in [21] for Gaussian vectors, with the difference that the results in this article apply to any kind of data mapped on a specific feature space. The regularization idea based on Gaussian convolutions is inspired on [22]. Finally, the use of Szegö's theorem exploiting the analogy between a probability density function and a power spectral density was also explored in [23] for KL divergence estimation by using autoregressive models for the densities.

This paper is organized as follows. Section II presents an overview of the SMI and finishes with a short outline of the proposed overall strategy. Then, Section III focuses on discrete sources and shows the fundamental link between the proposed surrogate and classical second-order statistics. Once the structure of the problem is unveiled, Section IV moves to analog sources along with the exposition of insightful tools for regularization and complexity reduction. The performance of the proposed estimators is illustrated by computer simulations in Section V, and Section VI summarizes the main conclusions of this work.

B. Notation

Column vectors: bold-faced lower case letters. Matrices: bold-faced upper case letters. $[\mathbf{A}]_{n,m}$: element at the n-th row and m-th column of matrix $\mathbf{A}$. $[\mathbf{a}]_n$: $[\mathbf{a}]_{n,1}$. $[\mathbf{a}]$: diagonal matrix with diagonal elements $[[\mathbf{a}]]_{n,n} = [\mathbf{a}]_n$. $(.)^T$: transpose. $(.)^H$: Hermitian transpose. $\mathrm{tr}(\mathbf{A})$: trace. $||.||$: Frobenius or Euclidean norm of a vector, matrix or function. $|.|$: absolute value of a complex number, or cardinality of a set. $\mathbf{A} \in \mathbb{R}^{N\times M}$: real matrix of dimension $N \times M$. $\mathbf{A} \in \mathbb{C}^{N\times M}$: complex matrix of dimension $N \times M$. $\mathbf{a} \in \mathbb{R}^N$: $\mathbf{a} \in \mathbb{R}^{N\times 1}$. $\mathbf{a} \in \mathbb{C}^N$: $\mathbf{a} \in \mathbb{C}^{N\times 1}$. $\mathbb{R}_+$: set of positive real numbers. $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$: $\mathbf{x}$ is a real Gaussian random vector of mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$. $\mathbf{x} \sim \mathcal{CN}(\boldsymbol{\mu}, \mathbf{R})$: $\mathbf{x}$ is a complex Gaussian random vector of mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$. $\mathbb{E}_p$: statistical expectation operator ($\mathbb{E}_p[f(x)] = \int f \mathrm{d}P$), where p is the mass function (or density function for analog variables), and P is the probability measure (or the cumulative distribution function for analog variables). $\hat{\mathbf{a}}$: an estimate of $\mathbf{a}$. $\langle x(l)\rangle_L = L^{-1}\sum_{l=1}^{L} x(l)$: L-th length sample mean operator. $\mathbf{I}_D$: $D \times D$ identity matrix. $\mathbf{0}_D$: $D \times 1$ vector containing all zeros. $\mathbf{1}_D$: $D \times 1$ vector containing all ones. 1_a: indicator function ($1_a = 1$ if a is true, and $1_a = 0$, otherwise). $\mathbf{A}^{1/2}$: Hermitian square root matrix of

the Hermitian matrix $\mathbf{A}$. $\mathbf{A}^{-1/2}$: Hermitian square root matrix of the Hermitian matrix $\mathbf{A}^{-1}$. $\mathbf{a}^{\alpha}$: element-wise power of a vector. $\alpha^{\mathbf{a}}$: element-wise power of a scalar. $\mathbf{a}^{T\alpha} = (\mathbf{a}^{\alpha})^{T}$. Toe($\mathbf{c}$): Toeplitz-Hermitian matrix constructed from its first column $\mathbf{c}$. $\odot$: Hadamard product. $*$: convolution operator. δ_{mn}: Kronecker delta. $\lceil x \rceil$: ceiling function.

II. Information-theoretic measures for data analytics

We begin by providing the rationale for identifying a surrogate that presents the desired properties for estimation. The derivation of the surrogate from its original measure, the Mutual Information (MI), is briefly addressed, and the relationship between both information measures is highlighted. After that, we outline the key ideas and the concrete goals of the paper. Through all the article, we will assume that $p_X(x)$, defined on a set $\mathcal{X}$, is either a mass function (for discrete sources) or a square-integrable density function (for continuous sources) associated with the random variable X.

A. Squared-loss mutual information as a surrogate

Consider two random sources X and Y defined on the sets $\mathcal{X}$ and $\mathcal{Y}$. The Pearson χ^2-divergence from $p_{XY}(x,y)$ to $p_X(x)p_Y(y)$, both defined on the product set $\mathcal{X}\times\mathcal{Y}$, is given by

$$D_{\chi^2}\left(p_{XY}||p_Xp_Y\right) = \mathbb{E}_{p_{XY}}\left[\frac{p_{XY}\left(x,y\right)}{p_X\left(x\right)p_Y\left(y\right)}\right] - 1 = I_s\left(X;Y\right), \tag{1}$$

where $I_s\left(X;Y\right)$ will be referred to as the Squared-loss Mutual Information (SMI), a term introduced in [29] for feature selection[1]. Alternatively, we can express the SMI as

$$\begin{aligned} I_s\left(X;Y\right) &= \mathbb{E}_{p_{XY}}\left[\left(\frac{p_{XY}\left(x,y\right) - p_X\left(x\right)p_Y\left(y\right)}{\sqrt{p_{XY}\left(x,y\right)p_X\left(x\right)p_Y\left(y\right)}}\right)^2\right] \\ &= \left\|\frac{p_{XY} - p_Xp_Y}{\sqrt{p_Xp_Y}}\right\|^2, \end{aligned} \tag{2}$$

which will be the definition used in this article. For a more in-depth exposition of (2), the reader is referred to Appendix A.

The SMI is also referred to as the mean-square contingency, a term characterized by Pearson in [31], and also studied by Rényi as a measure of dependence in [32]. The SMI can also be deduced from the called *normalized cross-covariance operator* (see [33], Eq. (9)), which corresponds to the kernel-free integral expression that is obtained from measuring dependence between two random variables through the Hilbert-Schmidt norm of a cross-covariance operator. This alternative way is of particular interest since it establishes a clear link between second-order statistics and the SMI, provided that the data is mapped onto certain feature spaces, as we will see in the next section.

[1]Although some researchers have defined the SMI as half of the magnitude of (2) [12], in this paper we will strictly define the SMI as it resolves from the χ^2-divergence. The half magnitude may be justified as closing the gap with the MI in the low dependence regime, but it then loses its physical significance.

In relation to other known measures of statistical dependence, the SMI is lower-bounded by

$$I_s\left(X;Y\right) \geq I_2\left(X;Y\right) \geq I\left(X;Y\right), \tag{3}$$

where $I_2(X;Y)$ is the second-order Rényi mutual information and $I(X;Y)$ is the Shannon MI. Since

$$I_2\left(X;Y\right) = \ln\left(1 + I_s\left(X;Y\right)\right), \tag{4}$$

the upper bound of $I_2(X;Y)$ is directly obtained from the fundamental logarithm inequality $\ln(1+x) \leq x$. Meanwhile, the strict inequality ($>$) of the MI is a consequence of the strict concavity of $\ln(.)$ for $p_{XY}(x,y) \neq p_X(x)p_Y(y)$. The equality (to zero) of both bounds is achieved only if $p_{XY}(x,y) = p_X(x)p_Y(y)$ [2]. This definition of $I_2(X;Y)$ is just the second-order Rényi divergence between the joint distribution and the product of the marginal distributions [24], which is in agreement with the definition in [25]. However, note that the characterization of the Rényi mutual information from the Rényi divergence can be accomplished differently [26]–[28].

Given that $I_2(X;Y)$ is an explicit and monotonic function of $I_s(X;Y)$ from (4), there is no practical difference between them in terms of computational complexity from data. For this reason and clarity, we will focus only on the SMI throughout this paper, having in mind that a tighter upper bound of $I_2(X;Y)$ can be obtained from $I_s(X;Y)$ via (4) if it were required for a particular application. It is also worth noting that, while the MI and the second-order Rényi MI satisfy the additivity property for independent (i.e. multiplicative) components, the SMI does not, given that the χ^2-divergence also does not satisfy it. Lastly, by following the general data processing inequality (see, for example, [34]), the SMI inherits the invariance property to nonlinear invertible transformations of the data from the second-order Rényi mutual information (Rényi divergence). This property constitutes a key point of the article, as we will rely on this property for estimation purposes.

B. Local approximation of mutual information

Once we have reviewed the surrogate that we propose for estimation, we are interested in further contextualizing its relationship with the MI. In this subsection, we review the property of the SMI of being a local approximation of the MI. This approximation becomes relevant for small values of dependence between random variables, which is a prominent case for applications in the context of Euclidean information theory [14].

Consider that $p_{XY}(x,y)$ and $p_X(x)p_Y(y)$ are close to each other, that is $p_X(x)p_Y(y) = p_{XY}(x,y) + \epsilon\Delta(x,y)$ for some small quantity ϵ, where $\Delta(x,y)$ is defined on the set $\mathcal{X}\times\mathcal{Y}$ and constrained to have null area. Without loss of generality, we will particularize the MI and SMI from the KL and χ^2 divergences, since they are the measures we are concerned about. Using the Taylor expansion of $\ln((1+\alpha)^{-1})$ up to the

[2]Note that these bounds follow from the χ^2, Rényi and KL divergence bounds, as can be seen in [24].

second order, i.e. $-\alpha+\alpha^2/2+O(\alpha^3)$, we can write the MI as follows:

$$\begin{aligned} I(X;Y) =& D\left(p_{XY}||p_{XY}+\epsilon\Delta\right) \\ =& \mathbb{E}_{p_{XY}}\left[\ln\frac{p_{XY}\left(x,y\right)}{p_{XY}\left(x,y\right)+\epsilon\Delta\left(x,y\right)}\right] \\ =& -\epsilon\mathbb{E}_{p_{XY}}\left[\frac{\Delta\left(x,y\right)}{p_{XY}\left(x,y\right)}\right] \\ &+\frac{1}{2}\epsilon^2\mathbb{E}_{p_{XY}}\left[\left(\frac{\Delta\left(x,y\right)}{p_{XY}\left(x,y\right)}\right)^2\right]+O\left(\epsilon^3\right). \end{aligned} \quad (5)$$

The first term is null since $\Delta(x,y)$ sums up zero, which implies that

$$I\left(X;Y\right)=\frac{1}{2}\epsilon^2\mathbb{E}_{p_{XY}}\left[\left(\frac{\Delta\left(x,y\right)}{p_{XY}\left(x,y\right)}\right)^2\right]+O\left(\epsilon^3\right). \quad (6)$$

Let us now examine the local behavior of the SMI. Using the Taylor expansion of $(1+\alpha)^{-1}$ up to the first order, i.e. $1-\alpha+O(\alpha^2)$, we can write the SMI in (2) as

$$\begin{aligned} &I_s\left(X;Y\right) \\ &=D_{\chi^2}\left(p_{XY}||p_{XY}+\epsilon\Delta\right) \\ &=\mathbb{E}_{p_{XY}}\left[\frac{\left(-\epsilon\Delta(x,y)\right)^2}{p_{XY}(x,y)\left(p_{XY}(x,y)+\epsilon\Delta(x,y)\right)}\right] \\ &=\epsilon^2\mathbb{E}_{p_{XY}}\left[\left(\frac{\Delta\left(x,y\right)}{p_{XY}\left(x,y\right)}\right)^2\left(1-\frac{\epsilon\Delta\left(x,y\right)}{p_{XY}\left(x,y\right)}+O\left(\epsilon^2\right)\right)\right] \\ &=\epsilon^2\mathbb{E}_{p_{XY}}\left[\left(\frac{\Delta\left(x,y\right)}{p_{XY}\left(x,y\right)}\right)^2\right]+O\left(\epsilon^2\right). \end{aligned} \quad (7)$$

From (6)&(7), the following fundamental result can be stated:

$$I\left(X;Y\right)=\frac{1}{2}I_s(X;Y)+O(\epsilon^3), \quad (8)$$

which means that half of the SMI, that is $\frac{1}{2}I_s$, constitutes a local approximation of the MI for close distributions. This observation is important because while I_s upper-bounds the MI, $\frac{1}{2}I_s$ is instead a local approximation, but not an upper bound.

The approximation of the SMI depicted in (8) is also studied as a relevant scenario in multiple related works. For instance, in [35] the approximation is analyzed in the problem of co-clustering contingency tables. On a more general note, we are particularly interested in the local approximations of the KL divergence and MI in [14] under the context of Linear Information Coupling (LIC) problems and Euclidean information theory. The main motivation is to translate information theory problems into linear algebra problems, thus avoiding computational and mathematical bottlenecks. Likewise, the relation between MI and SMI under local approximations has also been expressed in [15] (Eq. (61) and corresponding footnote), although more focused on providing an insightful measure of local information geometry. The derivation of the local approximation of the MI as a simpler problem where only linear algebra is involved is akin to the objective of this article, where the observation of the SMI as a second-order statistic opens the possibility of translating the problem of estimating information to a measure of the correlation between random variables, which is indeed a linear statistical dependency.

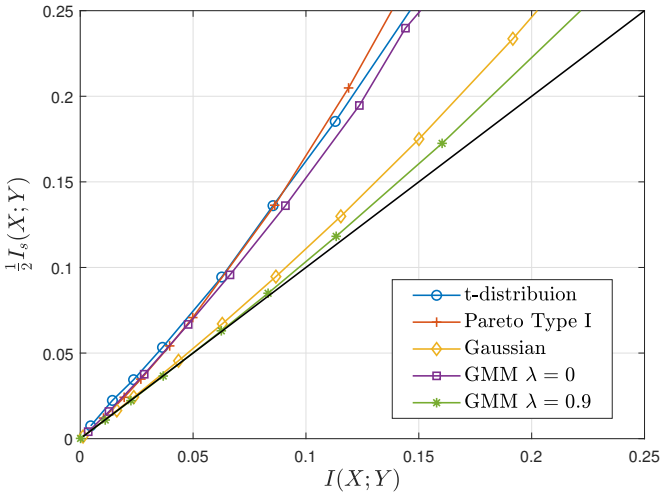

Fig. 1. Half of the SMI versus the MI for multiple distributions. The Student's t-distribution has $\nu=10$ degrees of freedom and the Pareto has location parameter $\theta=1$ for both marginal distributions. The GMM distribution and its parameters are detailed in Section V.

In order to illustrate the local approximation in (8), Figure 1 shows the closeness between half of the SMI and the MI of multiple distributions when they become close to independence. While the MI of the Student's t-distribution and the Pareto distribution is known (see [36] and [37], respectively), the SMI of both distributions is unknown. Nevertheless, their SMI can be estimated by a genie-aided estimator based on the empirical average of the function of interest under the knowledge of the marginal and joint distributions [2]. The Gaussian MI is known with $-0.5\ln(1-\rho^2)$ and its SMI can be obtained analytically and equals to $\rho^2/(1-\rho^2)$, where ρ is the Pearson correlation coefficient. The Gaussian Mixture Model (GMM) shown in the figure is based on the model shown in Section V, whose parameter λ determines different distributions with varying degrees of the local approximation in (8), as can be seen in the figure. For $\lambda=0$, the SMI is $\rho^4/(1-\rho^4)$, otherwise, both MI and SMI are computed with the genie-aided estimator.

C. Summary of the key idea

After the short review presented above, we can summarize the goal of this paper. The desired surrogate (2) has been exhibited, which has the following properties:

1) The surrogate is an upper bound of an information-theoretic measure of well-known operational meaning. By being an upper bound, we make sure that relevant information will not be lost by using surrogates for data analytics.
2) Half of the magnitude of the surrogate becomes close to a well-known information measure for the critical scenario of small dependence. It is then a meaningful measure in applications that require local approximations.

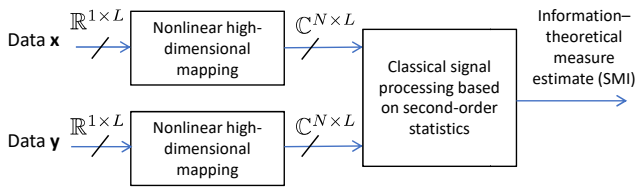

Fig. 2. Block diagram of the proposed data analytics strategy.

3) The surrogate can be expressed as a second-order moment, that is, as the expectation of a squared ratio of densities, and no logarithm is involved. The implication is that, by designing adequate preconditioning of the data, classical second-order analysis techniques should be enough for estimating information.

The purpose of what follows is to propose a universal mapping strategy from the data onto a high-dimensional feature space, such that the information can be extracted from that space by standard second-order signal processing techniques. The ultimate goal is to provide a rationale for the two-step data analytics strategy depicted in Figure 2. First, we analyze complex dependencies between two data sources by mapping L samples of the bivariate data onto a high-dimensional space[3]. The dimension N should be high enough to make sure that the maximum amount of complex associations potentially present in the data are captured, but it should be sufficiently small to provide reasonable computational complexity, as well as regularization capabilities. After that, the second stage is based on a second-order analysis that is focused on describing linear dependencies between sets of variables.

III. Discrete sources: second-order statistics on the simplex feature space

First, we focus our attention on discrete sources and the mapping onto the feature space. The objective is to determine the adequate preconditioning of the data and its restrictions, and to contextualize the rationale in terms of similar works. The results exhibited here will serve as a bridge for the pursuit of estimating information in the case of analog sources.

Consider that X and Y are discrete random variables with alphabets $\mathcal{X} = \{x_n\}_{n=1,2,...,N}$ and $\mathcal{Y} = \{y_m\}_{m=1,2,...,M}$, respectively. Let us define the marginal probability column vectors $\tilde{\boldsymbol{p}} \in \mathbb{R}_+^N$ and $\tilde{\boldsymbol{q}} \in \mathbb{R}_+^M$ with elements $[\tilde{\boldsymbol{p}}]_n = \Pr\{X = x_n\} = p_X(x_n)$ and $[\tilde{\boldsymbol{q}}]_m = \Pr\{Y = y_m\} = p_Y(y_m)$. Similarly, we define the joint probability matrix $\tilde{\mathbf{J}} \in \mathbb{R}_+^{N\times M}$ as $[\tilde{\mathbf{J}}]_{n,m} = \Pr\{X = x_n; Y = y_m\} = p_{XY}(x_n, y_m)$. Then, the SMI defined in (2) can be expressed as follows:

$$I_s(X;Y) = \sum_{n=1}^{N}\sum_{m=1}^{M}\left[\tilde{\mathbf{C}}\right]_{n,m}^2 = \mathrm{tr}\left(\tilde{\mathbf{C}}^T\tilde{\mathbf{C}}\right) = \left\|\tilde{\mathbf{C}}\right\|^2, \quad (9)$$

[3]In this article, we will define the feature space on the complex field given that the characteristic function will be used as the mapping function [38]. Nevertheless, a mapping into the space of real numbers can be used if necessary.

where

$$\tilde{\mathbf{C}} = [\tilde{\mathbf{p}}]^{-1/2}\left(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T\right)[\tilde{\mathbf{q}}]^{-1/2}. \quad (10)$$

Matrix $\tilde{\mathbf{C}} \in \mathbb{R}_+^{N\times M}$ in (10) will be referred to as *coherence matrix* due to its intimate link with the well-known CCA tool in statistical signal processing [39]. The form of matrix $\tilde{\mathbf{C}}$ is also encountered in the areas of information theory under the context of LIC problems [14] and Hirschfeld-Gebelein-Rényi (HGR) maximal correlation concept [32], [40]. To see these links, let us express

$$\tilde{\mathbf{C}} = \left(\mathbf{B} - \tilde{\mathbf{q}}^{1/2}\tilde{\mathbf{p}}^{T/2}\right)^T, \quad (11)$$

where

$$\mathbf{B} = [\tilde{\mathbf{q}}]^{-1/2}\,\tilde{\mathbf{J}}^T\,[\tilde{\mathbf{p}}]^{-1/2}. \quad (12)$$

We can write the joint probability mass function matrix as $\tilde{\mathbf{J}}^T = \mathbf{W}[\tilde{\mathbf{p}}]$, where $\mathbf{W} \in \mathbb{R}_+^{M\times N}$ is the channel transition matrix defined by the conditional probabilities $[\mathbf{W}]_{m,n} = \Pr\{Y = y_m | X = x_n\}$. Then we can express (12) as follows:

$$\mathbf{B} = [\tilde{\mathbf{q}}]^{-1/2}\,\mathbf{W}\,[\tilde{\mathbf{p}}]^{1/2}. \quad (13)$$

This matrix is called the *Divergence Transition Matrix* (DTM) of a discrete channel and it plays a fundamental role as a tool for translating information theory problems into linear algebra problems [14], [41]. Moreover, matrix $\tilde{\mathbf{C}}$ is the same as the *Canonical Dependence Matrix* (CDM) studied in [15] within the framework of modal decomposition of the joint probability mass function matrix. In this last case, the authors are interested in the universal features that define the underlying relationships among high-dimensional data. This paper strives to accomplish a similar rationale by translating a measure of information into a second-order statistics problem, thus with linear algebra in mind. However, while in [15] it is pursued as a matter of describing low-dimensional features, especially for the study of local approximations between distributions, in this work it is addressed as a mapping onto a high-dimensional space in order to extract those linear features, akin to kernel methods for measuring dependence [42].

Hereafter, we describe some important properties of the DTM and CDM matrices in the framework of this paper. In particular, in [14] it is shown that the maximum singular value of the DTM is $\sigma_1 = \sigma_{\max}(\mathbf{B}) = 1$, corresponding to right and left singular vectors $\tilde{\mathbf{p}}^{1/2}$ and $\tilde{\mathbf{q}}^{1/2}$, respectively. The second largest singular value of $\mathbf{B}$ corresponds to the largest singular value of matrix $\tilde{\mathbf{C}}$ [15], thus it is the second and subsequent singular values and vectors those who become fundamental in LIC problems. As a result from [14] and [15], the following proposition encompasses the implications of the singular values of $\tilde{\mathbf{C}}$ with the objective of estimating the SMI:

Proposition 1. *Let $\{\lambda_i\}_{i=1:\min(N,M)}$ be the singular values of the coherence matrix $\tilde{\mathbf{C}}$ in* (10)*. The largest singular value corresponds to the second largest singular value of matrix $\mathbf{B}$ from* (13)*, and the minimum singular value is zero. Therefore, the squared-loss mutual information in* (9) *is upper bounded by* $\min(N, M) - 1$.

Proof: Given that matrix $\mathbf{B}$ has singular values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{min(N,M)} \geq 0$, where σ_1 has $\tilde{\mathbf{p}}^{1/2}$ and $\tilde{\mathbf{q}}^{1/2}$ as

left and right singular vectors, it follows directly from (11) that $\lambda_{max} = \sigma_2$. From (9) and the definition of the Frobenius norm, it can be easily seen that the SMI requires the summation of all singular values of matrix $\tilde{\mathbf{C}}^T\tilde{\mathbf{C}}$, which correspond to the squared modulus of the singular values of $\tilde{\mathbf{C}}$. Since the maximum singular value is 1 and the minimum is 0, we obtain the stated upper bound on the SMI. ∎

The intuition behind this proposition is that $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{q}}$ define the probability simplex spaces constrained to the unit-sum $\sum_{n=1}^{N}[\tilde{\boldsymbol{p}}]_n = 1$ and $\sum_{m=1}^{M}[\tilde{\boldsymbol{q}}]_m = 1$. Hence, the contribution of one of the elements is lost (i.e. a singular value becomes 0) and only $N-1$ or $M-1$ are relevant. This notion will become meaningful in the next subsection, where the mapping onto the simplex contains the core idea of Theorem 1.

A. Relation to Canonical Correlation Analysis

We have seen that the unique expression of the SMI is featured in multiple areas of work and inherits some interesting properties. Most remarkably, no logarithm is involved in its definition, and $\tilde{\mathbf{C}}$ is related to the standard CCA method [43]. However, (9) does not contain covariance matrices as those required by the CCA. In the sequel, we unveil this relation by defining the required mapping in order to establish the full link between the SMI and the CCA. This link, jointly with the ideas provided in this subsection, is a particular derivation from the PICs between two discrete random variables provided in [17]. In this article, we will leverage these concepts to magnify the discussion of the SMI as a second-order statistic, and to focus on the structure and rank of the given mapping matrices, instead of the more general view of the SMI as a decomposition of the dependence among random variables.

First, let us express matrix $\tilde{\mathbf{C}}$ as a function of second-order statistics computed from the available data consisting of a sequence of L independent and identically distributed (i.i.d.) pairs $\{x(l), y(l)\} \in \mathcal{X}\times\mathcal{Y}$ for $l = 1, 2, \dots L$. Let $\hat{\tilde{\mathbf{p}}}$, $\hat{\tilde{\mathbf{q}}}$ and $\hat{\tilde{\mathbf{J}}}$ be estimates of the marginal and joint mass functions. From (9) we define the estimator of the SMI as

$$\hat{I}_s(X;Y) = \left\|\hat{\tilde{\mathbf{C}}}\right\|^2, \tag{14}$$

where $\hat{\tilde{\mathbf{C}}} = [\hat{\tilde{\mathbf{p}}}]^{-1/2}(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T)[\hat{\tilde{\mathbf{q}}}]^{-1/2}$ is the sample coherence matrix. To be specific, let us define the full-rank [4] data matrices $\mathbf{D}_x$ $(N\times L)$ and $\mathbf{D}_y$ $(M\times L)$ as follows:

$$[\mathbf{D}_x]_{n,l} = 1_{x(l)=x_n}, \qquad [\mathbf{D}_y]_{m,l} = 1_{y(l)=y_m}. \tag{15}$$

These data matrices are the result of a one-to-one mapping process from the elements of the sources to the canonical basis of dimension equal to the set cardinality. That is, column l of matrix $\mathbf{D}_x$ (or $\mathbf{D}_y$) is formed by $N-1$ (or $M-1$) zeros and a one in position n (or m). The mass function estimates required in (14) can be computed through first and second-order statistics as follows:

$$\hat{\tilde{\mathbf{p}}} = \frac{1}{L}\mathbf{D}_x\mathbf{1}, \qquad \hat{\tilde{\mathbf{q}}} = \frac{1}{L}\mathbf{D}_y\mathbf{1},$$

$$\left[\hat{\tilde{\mathbf{p}}}\right] = \frac{1}{L}\mathbf{D}_x\mathbf{D}_x^H, \qquad \left[\hat{\tilde{\mathbf{q}}}\right] = \frac{1}{L}\mathbf{D}_y\mathbf{D}_y^H,$$

$$\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T = \frac{1}{L}\mathbf{D}_x\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{D}_y^H, \tag{16}$$

where $\mathbf{P}_{\mathbf{1}}^{\perp} = \mathbf{I} - \mathbf{1}\mathbf{1}^T/L$ is the projection matrix onto the orthogonal space spanned by $\mathbf{1}$. As a result, the mass function estimates required in the computation of the SMI are just the two sample mean vectors, the two autocorrelation matrices and the cross-covariance matrix. The following lemma introduces a preliminary link with CCA:

Lemma 1. *Preliminary link SMI-CCA: Let $\mathbf{X}\in\mathbb{C}^{N\times L}$ and $\mathbf{Y}\in\mathbb{C}^{M\times L}$ be data matrices obtained as $\mathbf{X} = \mathbf{F}\mathbf{D}_x$ and $\mathbf{Y} = \mathbf{G}\mathbf{D}_y$, respectively, where $\mathbf{F}\in\mathbb{C}^{N\times N}$ and $\mathbf{G}\in\mathbb{C}^{M\times M}$ are full-rank mapping matrices. The estimated squared-loss mutual information based on a plug-in estimator is given by the Frobenius norm of a sample coherence matrix, that is:*

$$\left\|\hat{\mathbf{C}}\right\|^2 = \hat{I}_s(X;Y), \tag{17}$$

where

$$\hat{\mathbf{C}} = \hat{\mathbf{R}}_x^{-1/2}\hat{\mathbf{C}}_{xy}\hat{\mathbf{R}}_y^{-1/2}, \tag{18}$$

being $\hat{\mathbf{R}}_x = \mathbf{X}\mathbf{X}^H/L$ and $\hat{\mathbf{R}}_y = \mathbf{Y}\mathbf{Y}^H/L$ the sample autocorrelation matrices, and $\hat{\mathbf{C}}_{xy} = \mathbf{X}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{Y}^H/L$ the sample cross-covariance matrix.

Proof: Lemma 1 is a direct consequence of the SMI being invariant to nonsingular transformations. For this particular derivation, it can easily be proven given $||\hat{\mathbf{C}}||^2 = \text{tr}\left(\hat{\mathbf{C}}_{xy}\hat{\mathbf{R}}_y^{-1}\hat{\mathbf{C}}_{xy}^H\hat{\mathbf{R}}_x^{-1}\right)$, and using $\hat{\mathbf{R}}_x = \mathbf{X}\mathbf{X}^H/L = \mathbf{F}[\hat{\tilde{\mathbf{p}}}]\mathbf{F}^H$, $\hat{\mathbf{R}}_y = \mathbf{Y}\mathbf{Y}^H/L = \mathbf{G}[\hat{\tilde{\mathbf{q}}}]\mathbf{G}^H$ and $\hat{\mathbf{C}}_{xy} = \mathbf{X}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{Y}^H/L = \mathbf{F}(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T)\mathbf{G}^H$ from (16). Then, given that $\mathbf{F}$ and $\mathbf{G}$ are invertible, one can immediately obtain $||\hat{\mathbf{C}}||^2 = \text{tr}\left(\mathbf{F}(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T)[\hat{\tilde{\mathbf{q}}}]^{-1}(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T)^T[\hat{\tilde{\mathbf{p}}}]^{-1}\mathbf{F}^{-1}\right)$. By taking advantage of the circularity of the trace, we finally obtain

$$\left\|\hat{\mathbf{C}}\right\|^2 = \text{tr}\left(\left(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T\right)\left[\hat{\tilde{\mathbf{q}}}\right]^{-1}\left(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T\right)^T\left[\hat{\tilde{\mathbf{p}}}\right]^{-1}\right) = \left\|\hat{\tilde{\mathbf{C}}}\right\|^2. \tag{19}$$

∎

Lemma 1 sets the link between the SMI surrogate and second-order statistics. The implication is that we can estimate the SMI by mapping the events of sources X and Y onto the columns of matrices $\mathbf{F}$ and $\mathbf{G}$. These matrices can be seen as the *code-books* that contain all the possible column vectors $\{[\mathbf{F}]_{:,n}\}_{n=1,2,\dots,N}$ and $\{[\mathbf{G}]_{:,m}\}_{m=1,2,\dots,M}$ that compose the data matrices $\mathbf{X}$ and $\mathbf{Y}$, which are then used to construct the correlation and covariance matrices. It is important to highlight that $\mathbf{F}$ and $\mathbf{G}$ need to be full-rank matrices, thus a sufficient condition for equality in (17) is that $\mathbf{F} = \mathbf{I}_N$ and $\mathbf{G} = \mathbf{I}_M$, i.e. to map the data onto the orthonormal canonical basis. While the rank of the mapping matrices is not restrictive for discrete sources, this discussion will become relevant for analog sources.

Following the rationale, matrix $\hat{\mathbf{C}}$ in (18) is not (apparently) a coherence matrix as that required by CCA, because it is expressed in terms of autocorrelation matrices instead

[4]The data matrices are assumed full-rank for clarity, implying that L is sufficiently large such that $(x_n, y_m) \in \{x(l), y(l)\}_{l=1:L}$ for all $n = 1:N$ and $m = 1:M$. Note that $[\hat{\tilde{\mathbf{p}}}]$ and $[\hat{\tilde{\mathbf{q}}}]$ are therefore invertible under this assumption. The issue of rank-deficient data matrices will be specifically addressed later on.

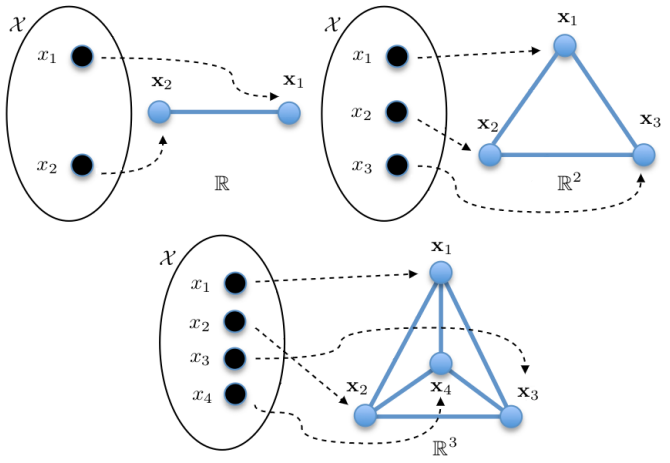

Fig. 3. Illustration of the mapping $\mathcal{X} \rightarrow \mathbb{R}^{|\mathcal{X}|-1}$ onto the $(|\mathcal{X}|-1)$-*simplex*.

of autocovariance matrices. However, the following theorem establishes the full link with CCA:

Theorem 1. *Full link SMI-CCA: Let* $\mathbf{X} \in \mathbb{C}^{N' \times L}$ $(N' < N)$ *and* $\mathbf{Y} \in \mathbb{C}^{M' \times L}$ $(M' < M)$ *be data matrices obtained as* $\mathbf{X} = \mathbf{F}\mathbf{D}_x$ *and* $\mathbf{Y} = \mathbf{G}\mathbf{D}_y$*, respectively, where* $\mathbf{F} \in \mathbb{C}^{N' \times N}$ *and* $\mathbf{G} \in \mathbb{C}^{M' \times M}$ *are full-rank mapping matrices. Let us define the small-size sample coherence matrix as*

$$\hat{\mathbf{K}}_{N',M'} = \hat{\mathbf{C}}_x^{-1/2} \hat{\mathbf{C}}_{xy} \hat{\mathbf{C}}_y^{-1/2}, \tag{20}$$

being $\hat{\mathbf{C}}_x = \mathbf{X}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{X}^H/L$ *and* $\hat{\mathbf{C}}_y = \mathbf{Y}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{Y}^H/L$ *the sample covariance matrices and* $\hat{\mathbf{C}}_{xy} = \mathbf{X}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{Y}^H/L$ *the sample cross-covariance matrix. Then:*

$$||\hat{\mathbf{K}}_{N',M'}||^2 = \hat{I}'_s\left(X;Y\right), \tag{21}$$

where

$$\hat{I}'_s\left(X;Y\right) \leq \hat{I}_s\left(X;Y\right). \tag{22}$$

In particular, a sufficient condition for the equality in (22) *is that* $N' = N-1$*,* $M' = M-1$ *and that the columns of* $\mathbf{F}$ *and* $\mathbf{G}$ *are given by the* $(N-1)$*-simplex and the* $(M-1)$*-simplex, respectively.*

Remark 1. In light of Theorem 1, we conclude that $\hat{\mathbf{K}}_{N-1,M-1}$ is just as valid as $\hat{\mathbf{C}}$ for estimating the SMI through their Frobenius norm, particularly for $N' = N-1$ and $M' = M-1$. For $N' < N$ and $M' < M$, the Moore-Penrose inverse is generally used to cope with the rank-deficient matrices [44].

Proof: See Appendix B. ∎

The implication of Lemma 1 and Theorem 1 is that, as $\hat{\mathbf{C}}$ (or $\hat{\mathbf{K}}_{N-1,M-1}$) is just the sample coherence matrix required in CCA, the squared-loss mutual information can be expressed just as the sum of the squared canonical correlations:

$$\hat{I}_s\left(X;Y\right) = \sum_{i=1}^{\min(N,M)-1} \hat{\lambda}_i^2\left(\hat{\mathbf{C}}\right). \tag{23}$$

The mapping onto the simplex is a direct consequence of Proposition 1 due to the simplex constraint, which serves as the interpretation of why the autocorrelation matrices are equally valid for the estimation of the SMI as the autocovariance matrices. Figure 3 illustrates the stated notion behind Theorem 1 and the simplex space: binary data (i.e. a discrete source with two possible outcomes) can be mapped to 1-dimensional points in the set $\{-1, 1\}$; ternary data (i.e. a discrete source with three possible outcomes) can be mapped to 2-dimensional points in the set $\{[1, 0], [-0.5, \sqrt{3}/2], [-0.5, -\sqrt{3}/2]\}$, and so on. Furthermore, this interplay between covariance and correlation for the estimation of the SMI may become advantageous due to the Toeplitz structure of the autocorrelation matrices, which arises in multiple applications in the signal processing field. Later on, this property is exploited in order to construct an estimator with reduced computational complexity.

In essence, Theorem 1 allows us to construct a desirable matrix form by mapping the events of the sources onto $\mathbf{F}$ and $\mathbf{G}$, and then we benefit from classical second-order statistics techniques. Note also that, since the coherence matrix is invariant under linear invertible transformations, the codebooks used for the SMI computation are irrelevant, provided that linearly independent vectors (columns of $\mathbf{F}$ and $\mathbf{G}$) are used. Otherwise, if the dimension of one or both of the spaces spanned after the mapping of X and Y are smaller than required (i.e. $N' < N-1$ and/or $M' < M-1$), the contribution of the smallest canonical correlations may be lost. The minimum dimension for the mapping of a source to vectors is then equal to the cardinality minus one. Moreover, the theorem also states implicitly that using higher dimensionality (i.e. $N' > N-1$ and/or $M' > M-1$) will yield a low-rank structure on $\hat{\mathbf{C}}_x$ and/or $\hat{\mathbf{C}}_y$. These ideas will take a fundamental role in the process of leveraging all these notions to the analog case: as the mapping requires an infinite dimension, the low-rank structure of the autocorrelation matrices will manifest and regularization is needed.

Regarding the related works in the literature that study the decomposition of the SMI, the reader is referred to [15] and [17]. Specifically, the consequences of Theorem 1 can also be deduced from [17]. However, the particular derivation in this paper provides a focused approach to the estimation of the SMI and the required mapping matrices. Specifically, the interplay between the marginal covariance and correlation matrices in terms of the link with CCA is not addressed in [17]. The final remark in (23) is also stated in [15] for the study of MI under weakly dependent variables. Furthermore, it is also shown that the HGR coefficient for discrete sources corresponds to the first singular value of $\hat{\mathbf{C}}$, while the SMI is given by the sum of the squares of all potentially nonzero singular values. Therefore, apart from the best single mapping that the HGR notion provides (i.e. the mapping that provides the largest correlation coefficient), the SMI looks as well to other mappings to canonical coordinates of the coherence, thus becoming more sensitive to complex hidden relationships between the observed data.

IV. Analog Sources: Second-Order Statistics on the Characteristic Feature Space

In the previous section we have shown that estimating the SMI via second-order statistics entails the mapping of events

onto a vectorial space spanning a minimum dimension equal to the source cardinality minus one. Nevertheless, analog sources require, in principle, a mapping to the function space to retain all the information. This key idea, informally stated in Cover's theorem on the separability of patterns [45][5], is well known in the field of machine learning. In particular, kernel methods have the ability (called *kernel trick* [46]) of implicitly using linear algebra on high (*infinite*) dimensional spaces without the necessity of explicitly visiting that space.

The objective of this section is to define a SMI estimator for analog sources based on the use of standard CCA as seen for discrete sources. To avoid the infinite-dimensional mapping, we will make use of a fixed dimension as a regularization of the problem from the beginning. This contrasts with kernel methods and with their implicit infinite mapping, where it is not so clear how to implement the inversion of matrices as those required by CCA (or the KCCA [47]), usually requiring strategies for decreasing complexity [46]. Besides, kernel methods also need to be regularized to avoid overfitting. In this sense, we propose an alternative based on an *explicit* mapping onto a space of *finite* dimension on the complex field, providing interpretability and computational complexity savings.

A. Dependence, correlation and characteristic function

To motivate the mapping, let us write the marginal and joint characteristic functions (CF) (defined as the Fourier transform of the PDFs with sign reversal in the complex exponential) of a pair of analog sources X and Y as follows:

$$\varphi_X(\omega_1) = \int p_X(x) e^{j\omega_1 x} \mathrm{d}x = \mathbb{E}_{p_X}\left[Z_1\right],$$
$$\varphi_Y(\omega_2) = \int p_Y(y) e^{j\omega_2 y} \mathrm{d}y = \mathbb{E}_{p_Y}\left[Z_2\right],$$
$$\varphi_{XY}(\omega_1,\omega_2) = \int p_{XY}(x,y) e^{j(\omega_1 x+\omega_2 y)} \mathrm{d}x\mathrm{d}y = \mathbb{E}_{p_{XY}}[Z_1 Z_2], \tag{24}$$

where $Z_1 = e^{j\omega_1 X}$ and $Z_2 = e^{j\omega_2 Y}$ are complex random variables obtained from X and Y through a nonlinear mapping. Clearly, if X and Y are independent, then $\varphi_{XY}(\omega_1,\omega_2) = \mathbb{E}_{p_X}\left[Z_1\right]\mathbb{E}_{p_Y}\left[Z_2\right] = \varphi_X(\omega_1)\varphi_Y(\omega_2)$ for all possible values of ω_1 and ω_2, implying that Z_1 and Z_2 are uncorrelated random variables. Note that the converse is also true: if Z_1 and Z_2 are uncorrelated for all possible values of ω_1 and ω_2, then X and Y are independent. This statement is given by the uniqueness property of the CF [38], and is a consequence of the bijective property of the Fourier transform. That is, if the condition $\varphi_{XY}(\omega_1,\omega_2) = \varphi_X(\omega_1)\varphi_Y(\omega_2)$ is true, then it implies that their probability distributions are also equal, i.e. $p_{XY}(x,y) = p_X(x)p_Y(y)$. This property of the CF has traditionally been used for the detection of independence between sources based on the difference between the joint and the product of marginal CFs [48], [49]. Moreover, the converse statement mentioned above guarantees that any kind of statistical dependence between X and Y will be "manifested" as correlation for some values of ω_1 and ω_2, which means that the set of complex exponential functions is not restrictive for the problem of independence detection via second-order statistics. In other words, an independence detector can also be formulated as a problem of detecting correlation by resorting to the CF, provided that a sufficient number of ω_1 and ω_2 values are explored [50].

For the case of estimation, we intend to provide an analysis of how many points of ω_1 and ω_2 need to be explored, i.e. which dimensionality is required by the mapping in order to estimate the SMI, and how small the separation between the explored points needs to be. For this purpose, we next propose a finite support for regularization and a uniform sampling of the CF (Sections IV-B&IV-C), which further yield an efficient estimation approach (Section IV-D).

B. Regularization through Gaussian convolutions

It is well known that the problem of estimating differential entropy and mutual information needs to be regularized [2]. In the sequel, we propose a regularization approach based on the properties of the CF. The core idea is the concept of Gaussian convolutions, which has been recently proposed in [22] in the framework of differential entropy estimation to achieve the parametric rate of convergence (w.r.t. the sample size) for distributions belonging to any nonparametric class. In the context of this paper, we also propose the use of Gaussian convolutions to regularize the estimation process. By deliberately contaminating the original random variables by means of a Gaussian CF with a known decay rate, we are then able to reduce the number of ω_1 and ω_2 points to be explored.

Consider that sources X and Y are contaminated by independent zero-mean additive Gaussian sources V_x and V_y with known smoothing variance σ^2 and PDF p_V:

$$x'(l) = x(l) + v_x(l), \quad y'(l) = y(l) + v_y(l). \tag{25}$$

The purpose is now to estimate the contaminated information between the virtual sources $x'(l)$ and $y'(l)$ using the data obtained from the actual sources $x(l)$ and $y(l)$. Note that the addition of the Gaussian noise is done by the construction of the estimator, therefore the original sources are still accessible. Since the PDF of the sum of independent random variables is the convolution of densities

$$p_{X'}(x) = p_X(x) * p_V(x), \quad p_{Y'}(y) = p_Y(y) * p_V(y), \tag{26}$$

then the CF is just the product of CFs of each term

$$\varphi_{X'}(\omega) = \varphi_X(\omega)\varphi_V(\omega), \quad \varphi_{Y'}(\omega) = \varphi_Y(\omega)\varphi_V(\omega), \tag{27}$$

where

$$\varphi_V(\omega) = e^{-\sigma^2 w^2/2} \tag{28}$$

is the CF of both V_x and V_y. This process can be cast as equivalently windowing the CF with a Gaussian function, where $\varphi_V(\omega)$ denotes the window function. The key point is that the Gaussian shape has an effective support, which allows

[5] "*A complex pattern-classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated*" (T. M. Cover).

focusing on a finite interval given by $|\omega| \leq \omega_{\max} = k\sigma^{-1}$. In particular, for any $\varepsilon > 0$, it holds that

$$|\varphi_{X'}(\omega)| < \varepsilon, \quad |\varphi_{Y'}(\omega)| < \varepsilon \tag{29}$$

for $\omega > \omega_{max}$. This bound can be immediately obtained by imposing $|\varphi_V(\omega)| < \varepsilon$ for $\omega > \omega_{max}$, and following the global bound $|\varphi_X(\omega)| \leq 1$ and $|\varphi_Y(\omega)| \leq 1$. Similarly, the joint CF is also bounded with

$$|\varphi_{X',Y'}(\omega_1, \omega_2)| < \varepsilon^2 \tag{30}$$

for $\omega_1 > \omega_{max}$ and $\omega_2 > \omega_{max}$, which is derived by following the separability of the CF for independent random variables

$$\varphi_{X',Y'}(\omega_1, \omega_2) = \varphi_{X,Y}(\omega_1, \omega_2)\, \varphi_V(\omega_1)\, \varphi_V(\omega_2). \tag{31}$$

Given $|\varphi_{X,Y}(\omega_1, \omega_2)| \leq 1$ and $|\varphi_V(\omega_1)| < \varepsilon$, $|\varphi_V(\omega_2)| < \varepsilon$ for $\omega_1 > \omega_{max}$, $\omega_2 > \omega_{max}$, we obtain (30). Then, we just need to choose a value of k that provides a considerably low value of ε given $\omega_{\max} = k\sigma^{-1}$. In this article, we propose the value of $k = 2.5$ as a trade-off between the effective support of the CF and the number of points to be explored. The consequence of the contamination is provided by the general data processing inequality for f-divergences (see [53] and references therein): the additive perturbation in both sources regularizes the problem by decreasing and bounding the amount of mutual information to be measured, yielding to a negative bias contribution to the estimators. This behavior will be verified later on with computer simulations.

It is important to note that the higher is σ^2, the stronger is the smoothing effect caused on the PDFs, and the smaller is the effective support of the contaminated CFs. This property exhibits an insightful duality with the classical spectral estimation problem [51]. Since the objective is to gradually reduce the values of the CF for increasing values of $|\omega|$, we may also refer to the window as the taper function [52]. In summary, the empirical CFs of the contaminated sources can be obtained by just tapering, or windowing, the sample mean estimators as follows:

$$\begin{aligned} \hat{\varphi}_{X'}(\omega) &= \left\langle e^{j\omega x(l)} \right\rangle_L \varphi_V(\omega), \\ \hat{\varphi}_{Y'}(\omega) &= \left\langle e^{j\omega y(l)} \right\rangle_L \varphi_V(\omega). \end{aligned} \tag{32}$$

The empirical estimation of the CF is known to be consistent for a reasonable wide class of probability distributions [54], and since the contaminating Gaussian CF is known, then (32) are also consistent estimators.

The next step is to determine the number of sampling points N of the CFs, which at the same time determines the dimension of the mapping. For this purpose, let us consider a uniform sampling in the ω domain with sampling period α. Since CFs and PDFs are Fourier pairs, the sampling of CFs implies a periodic extension of the PDFs, such that the implicit density of X becomes

$$p_{X'}(x) = \begin{cases} \sum_{k=-\infty}^{\infty} (p_X * p_V)\left(x - k\frac{2\pi}{\alpha}\right) & \frac{-\pi}{\alpha} \leq x \leq \frac{\pi}{\alpha} \\ 0 & \text{otherwise} \end{cases}, \tag{33}$$

and similarly for Y. The smaller is α, the smaller is the aliasing effect in (33). That is, we want to avoid as much as possible the overlapping of the replicas in (33). Therefore, the sampling period α can be roughly determined as the inverse of the expected dynamic range of the PDFs of the sources $\alpha = 1/(q\sigma_x)$, where σ_x is the standard deviation of the random variable X. In contrast to the choice of k, wherein the Gaussian CF has a well-behaved shape, q needs to be high enough to contain most of the PDF for a wide class of probability distributions. Hence, by following the Chebyshev's inequality [55], that is $\Pr(|X - \mu_x| \geq q\sigma_x) \leq q^{-2}$, where μ_x is the expected value of X, we propose to use $q = 6$. Note that, while lower values may become critical in terms of overlapping replicas in (33), higher values mainly contribute to an increase of N. Therefore, while q is chosen as a somewhat heuristic approach between the overlapping and the sampling size, it is reinforced by the Chebyshev's inequality.

Finally, we just need to equal the CF support $\omega_{\max} = k\sigma^{-1}$ to the highest value of the sampling $K\alpha$, where K is the value to be determined and tied to N. The number of sampling points of the CFs is then given by

$$N = 2\left\lceil \frac{\omega_{\max}}{\alpha} \right\rceil + 1 = 2\left\lceil kq\frac{\sigma_x}{\sigma} \right\rceil + 1, \tag{34}$$

where a CF support of $2\omega_{\max}$ and an odd value of N are imposed for a symmetric sampling around the origin. As a consequence, the additive Gaussian perturbation minimizes the effective support of the contaminated characteristic function (i.e. the dimension of the feature space) for a given smoothing variance σ^2, which further supports the rationale for using the tool of Gaussian convolutions as a natural regularization in the specific methodology explored in this paper. The interpretation of (34) is that of moving the problem to a finite parametric representation of the PDFs, which originally belongs to a non-parametric class. Then, as the implicit number of parameters of the problem becomes finite, the SMI estimation problem will turn out to be consistent.

C. Second-order statistics on the characteristic space and SMI estimate

Given the physical sense of the proposed regularization, we propose a uniform, symmetric and finite sampling of ω_1 and ω_2 to define the mappings $\phi_X(.) : \mathbb{R} \to \mathbb{C}^N$ and $\phi_Y(.) : \mathbb{R} \to \mathbb{C}^N$ as seen in [13], with

$$x \to \mathbf{x} = e^{j\alpha\mathbf{n}x}, \qquad y \to \mathbf{y} = e^{j\alpha\mathbf{n}y}, \tag{35}$$

respectively, where $\mathbf{n} \in \mathbb{Z}^{N\times 1}$ is a vector of integers defined as $\mathbf{n} = [-K, -K+1, \cdots, K]^T$ with $N = 2K+1$. To appreciate the rationale, note that if one lets $\alpha \to 0$ and $N \to \infty$ simultaneously in such a way that $N\alpha \to \infty$ as well (e.g. $\alpha = O(N^{-1/2})$), we are then mapping the sources onto asymptotically orthogonal vectors. If such condition is achieved, then the SMI estimate developed for discrete sources (based now on the CCA performed on the new spaces) will be asymptotically unbiased, according to Theorem 1. Finally, the feature space dimension is determined by (34) with $K = \lceil kq\sigma_x/\sigma \rceil$, which explains why using a finite dimension acts as a natural regularization of the problem.

Consider a sequence of L i.i.d. pairs $\{x(l), y(l)\} \in \mathbb{R}^2$ for $l = 1, 2, \ldots, L$. Using the mapping defined in (35), we

obtain the pair of vector sequences $\{\mathbf{x}(l), \mathbf{y}(l)\} \in \mathbb{C}^{N\times 2}$ in the feature space and construct the data matrices $\mathbf{X} \in \mathbb{C}^{N\times L}$ and $\mathbf{Y} \in \mathbb{C}^{N\times L}$ as follows:

$$[\mathbf{X}]_{:,l} = \mathbf{x}(l), \qquad [\mathbf{Y}]_{:,l} = \mathbf{y}(l). \tag{36}$$

On the one hand, the cross-covariance matrix will be defined following (32) with

$$\hat{\mathbf{C}}_{x'y'} = \left\langle e^{j\alpha \mathbf{n} x(l)} e^{-j\alpha \mathbf{n}^T y(l)} \right\rangle_L \odot \left(\mathbf{w}\mathbf{w}^T\right) - \hat{\mathbf{p}}\hat{\mathbf{q}}^H, \tag{37}$$

where the weighted first-order statistics are

$$\hat{\mathbf{p}} = \left\langle e^{j\alpha \mathbf{n} x(l)} \right\rangle_L \odot \mathbf{w}, \quad \hat{\mathbf{q}} = \left\langle e^{j\alpha \mathbf{n} y(l)} \right\rangle_L \odot \mathbf{w}, \tag{38}$$

and the symmetric tapering vector is defined as

$$[\mathbf{w}]_n = \varphi_V\left((n-K)\,\alpha\right) = e^{-\sigma^2\alpha^2(n-K)^2/2} \tag{39}$$

for $n = 0, 1, \ldots, N-1$. Note that $\mathbf{w}$ follows from the concept of Gaussian convolutions in (28), therefore the matrix in (37) refers to the contaminated sources X' and Y'.

On the other hand, the elements of the sample autocorrelation matrices can be expressed as

$$\begin{aligned}\left[\hat{\mathbf{R}}_{x'}\right]_{n,m} &= \left\langle e^{j\alpha(n-m)x(l)} \right\rangle_L \varphi_V\left(\alpha\,(n-m)\right), \\ \left[\hat{\mathbf{R}}_{y'}\right]_{n,m} &= \left\langle e^{j\alpha(n-m)y(l)} \right\rangle_L \varphi_V\left(\alpha\,(n-m)\right)\end{aligned} \tag{40}$$

for $n, m = 0, 1, \ldots, 2K$, which endow them with a Toeplitz structure. As a result, we can construct them as follows:

$$\hat{\mathbf{R}}_{x'} = \text{Toe}\left(\hat{\mathbf{p}}_a\right), \quad \hat{\mathbf{R}}_{y'} = \text{Toe}\left(\hat{\mathbf{q}}_a\right), \tag{41}$$

where $\hat{\mathbf{p}}_a$ and $\hat{\mathbf{q}}_a$ are defined as the extended weighted first-order statistics,

$$\hat{\mathbf{p}}_a = \left\langle e^{j\alpha \mathbf{n}_a x(l)} \right\rangle_L \odot \mathbf{w}_a, \quad \hat{\mathbf{q}}_a = \left\langle e^{j\alpha \mathbf{n}_a y(l)} \right\rangle_L \odot \mathbf{w}_a, \tag{42}$$

with $\mathbf{n}_a = [0, 1, \cdots, N-1]^T$, and the asymmetric tapering vector

$$[\mathbf{w}_a]_n = \varphi_V(n\alpha) = e^{-\sigma^2\alpha^2 n^2/2} \tag{43}$$

for $n = 0, 1, \ldots, N-1$. Similarly with (37), the correlation matrices also refer from the contaminated sources X' and Y' due to the effect of the tapering vector.

Finally, once the matrices involved in the estimation of the SMI for analog sources are defined, from Remark 1 we have that

$$\hat{I}_s(X;Y) = \left\|\hat{\mathbf{C}}\right\|^2, \tag{44}$$

with

$$\hat{\mathbf{C}} = \hat{\mathbf{R}}_{x'}^{-1/2} \hat{\mathbf{C}}_{x'y'} \hat{\mathbf{R}}_{y'}^{-1/2}. \tag{45}$$

Note that we have translated the problem of estimating the SMI to the problem of estimating the cross-covariance and autocorrelation matrices, which are known to be consistent for i.i.d. data. This is a desired property inherited from the case of discrete sources. Given that the empirical CFs are consistent, as well as the aforementioned matrices, the convergence of the estimator (44) is determined only by the parameters N and α. For a fixed α, the rationale for choosing σ^2, and therefore N following (34), will be detailed in Section V.

As a final remark, the regularization technique proposed above differs from the classical regularization technique used in the KCCA based on diagonal loading of autocorrelation matrices [18]. Although both techniques succeed in solving the rank-deficient issue, the proposed regularization provides a physical interpretation of the overall effect on the final estimate.

D. Large feature space dimension regime approximation

The Toeplitz structure of $\hat{\mathbf{R}}_{x'}$ and $\hat{\mathbf{R}}_{y'}$ can be further exploited for the computation of the inverses in (45). Szegö's theorem (see [56], [57]) establishes that a Toeplitz matrix is asymptotically diagonalizable by the unitary Fourier matrix, and its eigenvalues asymptotically behave like samples of the Fourier transform of its entries for an increasing matrix size. The most general and relaxed assumption that guarantees the behavior stated in Szegö's theorem is that the columns of the matrices are square-integrable for $N \to \infty$. This condition is clearly ensured by the tapering operation in (42): as the Gaussian taper in (43) is square-integrable for any $\sigma^2 > 0$ and the sample CFs are upper-bounded, that is $|\left\langle e^{j\alpha n x(l)} \right\rangle_L| \leq 1$ and $|\left\langle e^{j\alpha n y(l)} \right\rangle_L| \leq 1$, then the sample vectors $\hat{\mathbf{p}}_a$ and $\hat{\mathbf{q}}_a$ become square-integrable for $N \to \infty$. This fact motivates a frequency-domain tool to reduce complexity by leveraging the approximate diagonalization of the involved Toeplitz matrices after a Fourier transform.

The following lemma sets the required theoretical framework:

Lemma 2. *Let $t_n \in \mathbb{C}$ be an Hermitian sequence such that $t_0 = 1$ and $\lim_{N\to\infty} \sum_{n=0}^{N-1} |t_n|^2 < \infty$. Let us define vector $\mathbf{t} \in \mathbb{C}^N$ and Hermitian-Toeplitz matrix $\mathbf{T} \in \mathbb{C}^{N\times N}$ as $[\mathbf{t}]_n = t_n$ and $\mathbf{T} = \text{Toe}(\mathbf{t})$, respectively. Let $\mathbf{U} \in \mathbb{C}^{N\times N}$ be the unitary Fourier matrix and $\mathbf{H} = \mathbf{U}\mathbf{T}\mathbf{U}^H$. Then*

$$\lim_{N\to\infty} \left\{ [\mathbf{H}]_{n,m} - \left(2\sqrt{N}\,Re\left(\left[\mathbf{U}^H(\mathbf{t}\odot\mathbf{v})\right]_n\right) - 1\right)\delta_{nm} \right\} = 0 \tag{46}$$

for $n, m = 0, 1, \ldots, N-1$, where $\mathbf{v}$ is a unilateral triangular window with elements $[\mathbf{v}]_n = 1 - n/N$.

Proof: See [57] for a detailed proof concerning the limit behavior. In addition, in (46) we have used that the Fourier transform of a Hermitian sequence g_n can be written as $\sum_{n=-(N-1)}^{(N-1)} g_n e^{-j2\pi wn} = g_0 + 2\text{Re}\left(\sum_{n=1}^{(N-1)} g_n e^{-j2\pi wn}\right) = 2\text{Re}\left(\sum_{n=0}^{(N-1)} g_n e^{-j2\pi wn}\right) - g_0$, and that $g_n = t_n(1 - n/N)$ with $g_0 = 1$. ∎

In light of Lemma 2, we can now build an approximate and computationally efficient SMI estimator for the large dimension regime.

Theorem 2. *Let $\hat{\mathbf{C}}_{x'y'} \in \mathbb{C}^{N\times N}$ be the estimated cross-covariance matrix from (37), $\hat{\mathbf{p}}_a \in \mathbb{C}^N$ and $\hat{\mathbf{q}}_a \in \mathbb{C}^N$ the estimated expected values of the mapped data from (42), $\mathbf{U} \in \mathbb{C}^{N\times N}$ the unitary Fourier matrix and $\mathbf{v} \in \mathbb{C}^N$ a unilateral triangular window. Then, the reduced computational complexity estimator of the squared-loss mutual information is as follows:*

$$\hat{I}_{as}(X;Y) = \left\| \left[\hat{\mathbf{p}}'\right]^{-1/2} \mathbf{U}\hat{\mathbf{C}}_{x'y'}\mathbf{U}^H \left[\hat{\mathbf{q}}'\right]^{-1/2} \right\|^2, \tag{47}$$

where

$$\hat{\mathbf{p}}' = 2\sqrt{N}Re\left(\mathbf{U}^H\left(\hat{\mathbf{p}}_a \odot \mathbf{v}\right)\right) - \mathbf{1},$$
$$\hat{\mathbf{q}}' = 2\sqrt{N}Re\left(\mathbf{U}^H\left(\hat{\mathbf{q}}_a \odot \mathbf{v}\right)\right) - \mathbf{1}. \quad (48)$$

Proof: From (44), and given that the Frobenius norm is invariant under unitary transforms, the SMI estimator can be written as

$$\hat{I}_s\left(X;Y\right) = \left\| (\mathbf{U}\hat{\mathbf{R}}_{x'}\mathbf{U}^H)^{-1/2}\mathbf{U}\hat{\mathbf{C}}_{x'y'}\mathbf{U}^H(\mathbf{U}\hat{\mathbf{R}}_{y'}\mathbf{U}^H)^{-1/2} \right\|^2. \quad (49)$$

Then, following Lemma 2, the matrices $\mathbf{U}\hat{\mathbf{R}}_{x'}\mathbf{U}^H$ and $\mathbf{U}\hat{\mathbf{R}}_{y'}\mathbf{U}^H$ are approximated by $[\hat{\mathbf{p}}']$ and $[\hat{\mathbf{q}}']$, respectively, given that $[\hat{\mathbf{p}}_a]_0 = [\hat{\mathbf{q}}_a]_0 = 1$. ∎

The main advantage of the proposed approximate estimator in (47) is that the inverses of $\hat{\mathbf{R}}_{x'}$ and $\hat{\mathbf{R}}_{y'}$ are avoided and only element-wise inverses are required, but a high value of N is needed to cope with the limit behavior. However, since the computational complexity of the estimator increases with N, the approximate estimator becomes more relevant if a higher dimensionality is needed, thus reducing complexity.

V. Numerical results

In this section, the performance of the proposed estimators and the impact of their free parameters are evaluated by means of Monte Carlo simulations. Unless otherwise stated, we will show the performance of the estimator with i.i.d. data modeled as:

$$x\left(l\right) = h_x\left(l\right)\left(-\sqrt{\lambda} + z\left(l\right)\sqrt{1-\lambda}\right)\sqrt{\rho} + w_x\left(l\right)\sqrt{1-\rho},$$
$$y\left(l\right) = h_y\left(l\right)\left(\sqrt{\lambda} + z\left(l\right)\sqrt{1-\lambda}\right)\sqrt{\rho} + w_y\left(l\right)\sqrt{1-\rho}, \quad (50)$$

where $z, w_x, w_y \sim \mathcal{N}\left(0,1\right)$ are also i.i.d. random variables. The variables h_x, h_y are also independent from each other and take values $h_x, h_y \in \{-1, 1\}$ with equal probability. The joint distribution is then expressed as a GMM, which is exhibited in Appendix C. While the correlation coefficient $\rho \in [0, 1)$ determines the degree of dependence, the parameter $\lambda \in [0, 1]$ determines multiple distributions with different difficulty levels of SMI estimation. Specifically, for $\lambda = 0$ the distribution follows from a GMM with just two zero-mean Gaussian components of ρ and $-\rho$ correlation coefficients, and is a distribution that has been used previously as a model for detecting and measuring dependence (see [58], [13]). This particular case has the advantage of a known SMI under any value of the smoothing variance from (28) with

$$I_s\left(X,Y\right)_{\sigma^2,\lambda=0} = \frac{\rho^4}{\left(1+\sigma^2\right)^4 - \rho^4}, \quad (51)$$

which allows us to characterize the expected value of the SMI under a Gaussian regularization.

Overall, the usefulness of the model lies in the fact that $\mathbb{E}_{p_{XY}}[XY] = 0$ for any value of λ. This property is particularly interesting given that the proposed method is based on measuring correlation in the feature space, thus forcing the estimators to discover dependence from originally uncorrelated data. For simplicity, the model has also been calibrated such that $\mathbb{E}_{p_X}[X^2] = \mathbb{E}_{p_Y}[Y^2] = 1$. Figure 4 shows the proposed GMM for $\lambda = 0$ and $\lambda = 0.9$ in comparison with the other distributions also illustrated in Figure 1. We use Student's t-distribution and the Pareto as examples of distributions with long tails and, for the Pareto specifically, a unilateral exponential distribution. For illustration purposes, all distributions parameters have been calibrated to have, approximately, the same value of SMI.

In terms of parameter choices, smoothing variance σ^2 and dimensionality of the mapping N, the rationale is the following. For the selection of σ^2, the estimator uses the classical Silverman's rule [59], [60] derived in the context of nonparametric kernel density functionals estimation, which is known to provide consistent results for small dimensional data [4]. According to this rule, the perturbation variance is set to monotonically decrease with L as

$$\sigma^2 = pL^{-2/5}, \quad (52)$$

being p the new free parameter. The rationale behind the choice of σ^2 is that of determining the convergence rate of the estimator. This relation between data size and perturbation variance can also be encountered in the context of spectral density estimation after minimizing the Mean Squared Error (MSE) with respect to the taper bandwidth [61], recalling the resemblance between the perturbation based on Gaussian convolutions and the spectral estimation problem. For clarity, the rationale for using this rule is shown in Appendix D. For selecting p it is enough to choose a sufficiently small value, but it can be increased (thus N decreased) if small values of SMI are expected since these are more robust to higher contamination σ^2 (for example, see (51)). Then, N is computed following (34) with the values of $q = 6$ and $k = 2.5$ elaborated and provided in Section IV-B.

Figure 5 shows the mean of the proposed SMI estimators as a function of small ($I_s\left(X,Y\right) \in [0, 0.1]$) and moderate ($I_s\left(X,Y\right) \in (0.1, 1]$) values of true SMI for different values of the parameter p and data size L. The objective is twofold: to illustrate the inherent negative bias introduced by regularizing, which increases with σ^2 (as N decreases) and is provided by the genie-aided estimator, and to show that the estimator becomes asymptotically close from above to the contaminated SMI value as $L \to \infty$. In short, the small dependence regime is the data limited regime, and the strong dependence regime is the dimension limited regime, thus a smaller p is required as the SMI increases. Moreover, the residual biases associated with estimates of theoretically null squared canonical correlations cause a ground level for sufficiently small values of p. In order to compensate this behavior, a reduced bias estimator $\hat{I}_s\left(X,Y\right) - \hat{I}_s\left(X,Y_{\text{ind}}\right)$ is also shown, where Y_{ind} is i.i.d. data like Y and obtained by circularly shifting the data sequence associated to Y by j positions with $j \neq 0$ and $j \neq L$, thus improving the impact on the overall bias at the small data regime regardless of the kind of data statistics.

In order to assess the performance of the estimator for multiple classes of distributions, Figure 6 shows the expected value of the estimator for a varying value of data size L, showing the same distributions depicted in Figure 4. In this case, we have fixed the value of σ^2 and N for any L to

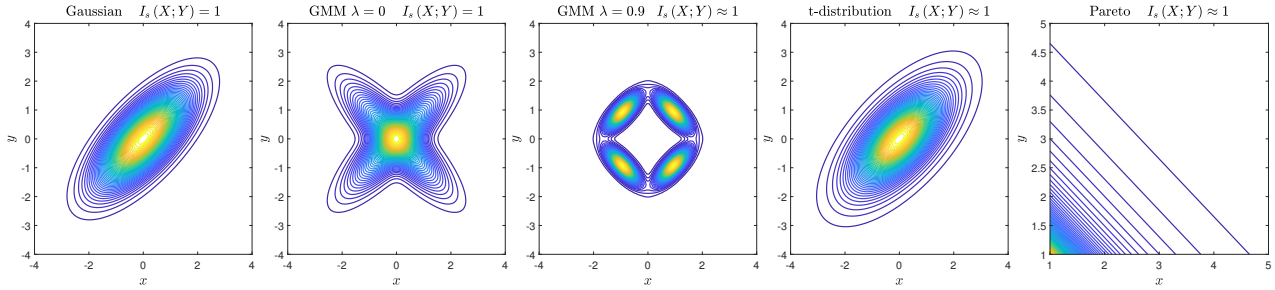

Fig. 4. Contour plots of the Gaussian distribution, the proposed GMM for two values of λ, Student's t-distribution and the Pareto Type I distribution.

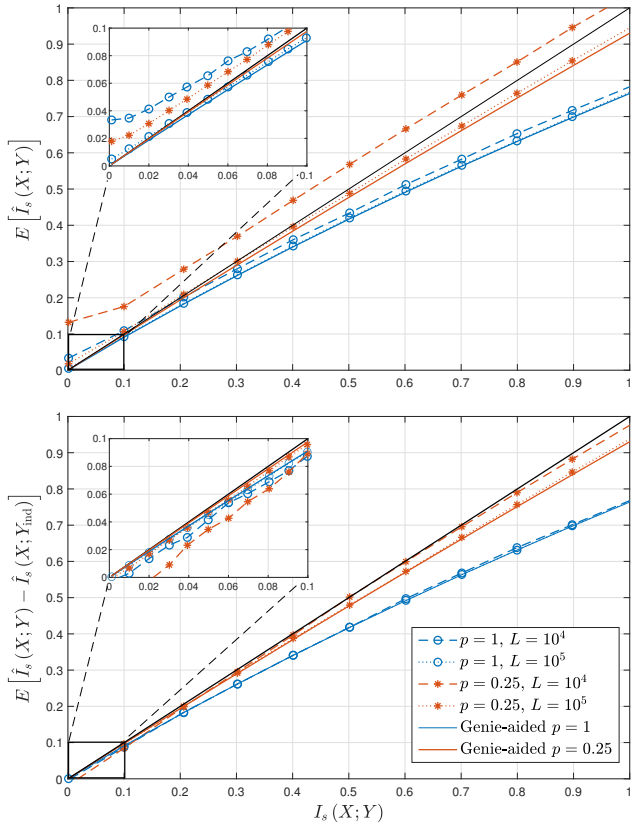

Fig. 5. Mean of the estimated SMI (up) and reduced-bias estimators (down) as a function of the true SMI for $p = 1$ ($\sigma^2 = 10^{-2}$, $N = 301$) and $p = 0.25$ ($\sigma^2 = 2.5 \times 10^{-3}$, $N = 601$), showing the role of σ^2 and L, with $N = 2\lceil qk/\sigma \rceil + 1$, $q = 6$ and $k = 2.5$. The data is modeled following the GMM with $\lambda = 0.9$.

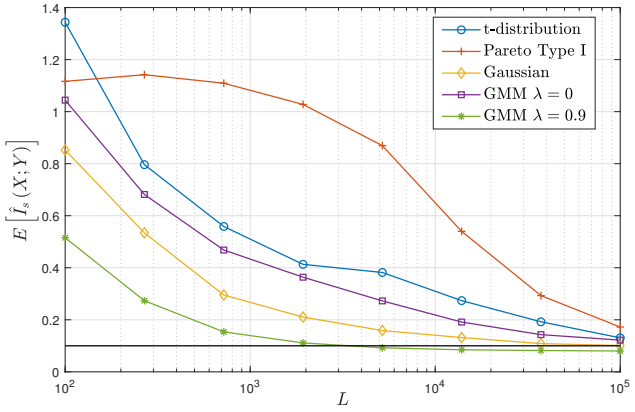

Fig. 6. Mean of the estimated SMI for multiple distributions as a function of the data size L, and SMI value of 0.1, displayed in the black line. The parameters are $p = 2.5$, $q = 6$ and $k = 2.5$, which yield $\sigma^2 = 2.5 \times 10^{-2}$ and $N = 191$ with a fixed objective value of $L = 10^5$.

properly show the convergence of the estimator, specified in the figures caption. Although the rate of convergence may vary between different distributions, it is shown that all of them tend to the contaminated value of SMI. This is especially clear for the GMM with $\lambda = 0.9$, which attains a faster convergence rate and the negative bias is manifested. On the contrary, the Pareto distribution is specifically challenging since its CF is not absolutely continuous for all values of ω, which hinders the estimation through the proposed feature map.

Figure 7 depicts the Normalized Mean Squared Error (NMSE) versus L of the proposed estimator $\hat{I}_s(X,Y)$ along with its reduced bias version $\hat{I}_s(X,Y) - \hat{I}_s(X,Y_{\text{ind}})$. In this figure, the estimator computes an adequate value of σ^2 and N following the rationale from Appendix D. That is, for each value of L an intermediate perturbation variance and dimension are computed in order to attain the desired convergence rate with respect to the MSE. For completeness, three more estimators are shown: the least-squares mutual information estimator (LSMI) [29], the estimator based on the Adaptive Partitioning (AP) of the observation space [9], and the one based on Kernel Density Estimation (KDE) [7]. While the LSMI is an explicit estimator of the SMI, both AP and KDE are plug-in density estimators that have been adjusted to estimate the SMI. The LSMI parameters are chosen by a cross-validation procedure as in [29], while the kernel bandwidth of the KDE is chosen following the rule provided in [60]. The figure also shows the GMM distribution for two values of λ, following Figure 4, and therefore the performance is assessed for different scenarios. Generally speaking, it can be seen that the proposed estimator and rationale behind parameter choices are effective for increasing values of L. In terms of comparative performance, the proposed method generally attains similar (or better) NMSE values than the other shown estimators for low values of L, except for the particularly difficult case $\lambda = 0$ and $I_s = 1$. The main advantage of our estimator is that, contrary to the other methods, its computational complexity scales with the dimension N rather than with the sample size L. This is especially true for the

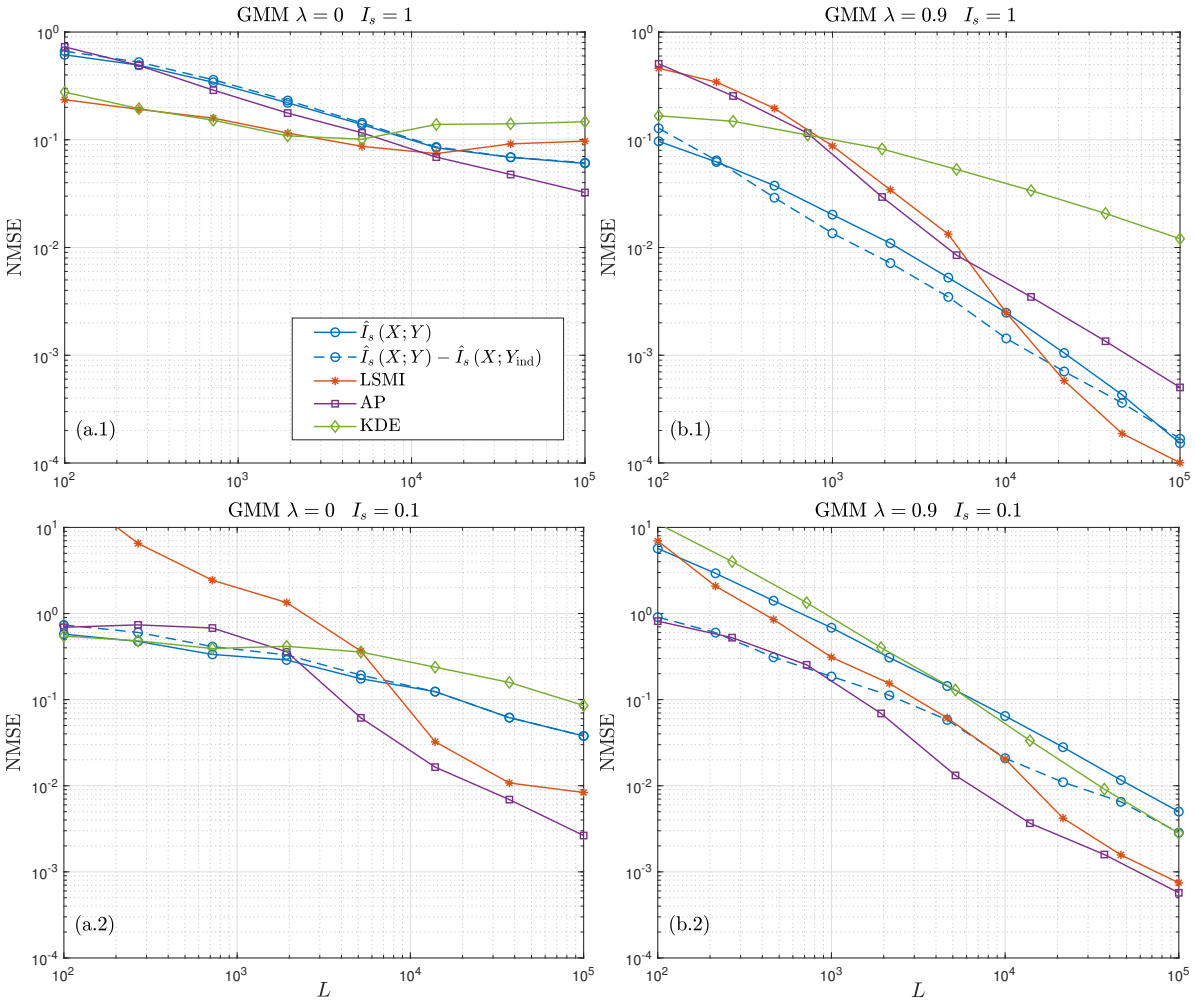

Fig. 7. NMSE of the estimated SMI as a function of data size L for $q = 6$, $k = 2.5$, $\sigma^2 = pL^{-2/5}$ and $N = 2\lceil qk/\sigma \rceil + 1$. Parameters of choice: $(a.1)$: $p = 2.5$, $(a.2)$: $p = 5$, $(b.1)$: $p = 0.1$, $(b.2)$: $p = 0.25$.

KDE and LSMI estimators, whose computational complexity is $\mathcal{O}\left(L^2\right)$ [62], and due to the cross-validation step, respectively, while the proposed approach stays at $\mathcal{O}\left(N^2\right)$. On the other hand, the complexity of the AP estimator depends on the partitioning algorithm [63], although it still scales with L. The advantage becomes more appealing for the case $N \ll L$, where the proposed method is less computationally intensive at the cost of performance, thus presenting a trade-off between complexity and accuracy. The effectiveness of the reduced bias estimator is also shown, especially for the case $\lambda = 0.9$ given that the dimensionality N is higher, thus contributing to a higher ground value. For the case $\lambda = 0$, and lower values of N, the ground level can be neglected and the cost of the reduced bias estimator is only a slightly higher variance. Finally, it can be observed the choice of p: while easier scenarios require very small values of p (small values of contamination), more difficult scenarios need higher values of p, thus a stronger regularization.

Finally, the performance of the approximate frequency-domain estimator described in Section IV-D is shown in Figure 8 in terms of the bias. It can be seen that, as the dimension increases, the performance of the approximate estimator converges to that of the original estimator (provided that a nonzero smoothing variance is used) with the advantage of a significantly reduced computational load. Note that the greater is the smoothing variance, the faster is the convergence rate of the frequency-domain estimator to the original performance, at the expense of an increased negative bias. Moreover, the estimator without regularization is also shown, which corresponds to $p = 0$. In this case, it can be seen that it diverges from the true value of SMI for an increasing value of N, which demonstrates the need of regularizing the proposed estimator.

VI. Conclusion

In this paper, we have derived an estimator of the degree of dependence between a pair of i.i.d. data based solely on second-order statistics computed after mapping the data onto a finite-dimensional complex space. The use of second-order statistics is possible as a result of selecting a surrogate of mutual information that is a quadratic measure of dependence.

In particular, it is shown that the squared-loss mutual information used in the field of machine learning corresponds to second-order statistics based on the Frobenius norm of a coherence matrix, which is known to be directly linked with the

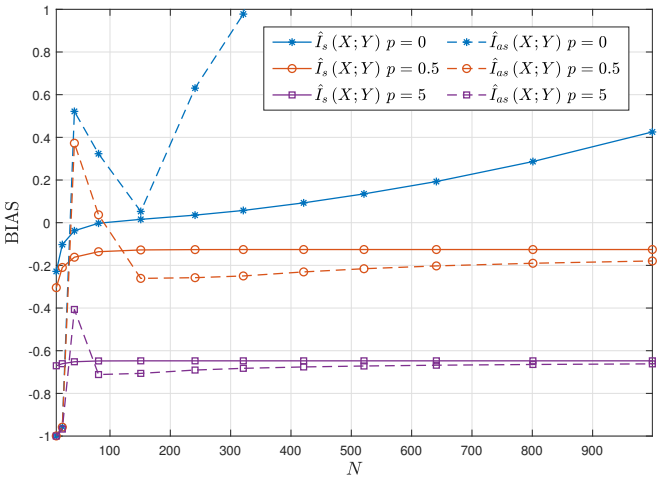

Fig. 8. Bias of the SMI estimator and the approximate SMI estimator as a function of feature space dimension N, for SMI $= 1$, $q = 6$, $L = 10^5$, and $\sigma^2 = p/\left(L^{2/5}\right)$. The data is modeled following the GMM with $\lambda = 0.9$.

standard CCA tool. Moreover, the selected surrogate has the property of upper-bounding the Shannon mutual information, and it behaves as a local approximation of twice the mutual information in the small dependence regime. The theoretical background is contextualized within the fields of information theory and signal processing, where some connections with well-known concepts in the literature have emerged, such as the locally optimal detector of correlation for Gaussian data, the linear information coupling problems, and the spectral density estimation problems.

While in the case of discrete data it suffices to map the data onto the $(N-1)$-simplex, for analog data the natural feature space is based on steering vectors and its dimension can be selected as a regularization parameter of the problem, trading off performance (bias) and complexity. The main advantage of avoiding the dual form as in kernel methods is that the estimators become linearly scalable with respect to the data size, and that the free parameters can be selected with physical meaning related to the expected dynamic range and expected smoothing degree of the true densities. However, the resulting rule for deciding the free parameters is still vulnerable to different values of the true squared-loss mutual information, especially for distributions whose characteristic function is not absolutely continuous.

Finally, some pending issues are left for future work, such as the extension of the estimator to the case of data with memory, as proposed for instance in [64], and a data-dependent dimensionality reduction strategy prior to CCA, for which some preliminary results based on the *minimum description length* principle have recently been provided in [65].

VII. Appendices

Appendix A: Derivation of (2).

Defining the joint distribution and the product of the marginal distributions on the product set $\mathcal{X} \times \mathcal{Y}$ and letting P_{XY} as the probability measure, we have

$$\begin{aligned} I_s\left(p_{XY}||p_X p_Y\right) &= \int\int \frac{p_{XY}}{p_X p_Y}\mathrm{d}P_{XY} - 1 \\ &= \int\int \left(\frac{p_{XY}}{p_X p_Y} - 2\right)\mathrm{d}P_{XY} + 1 \\ &= \int\int \frac{p_{XY}^2 - 2p_{XY}p_X p_Y}{p_{XY}p_X p_Y}\mathrm{d}P_{XY} \\ &\quad + \int\int \frac{p_X^2 p_Y^2}{p_{XY}p_X p_Y}\mathrm{d}P_{XY} \\ &= \int\int \frac{p_{XY}^2 - 2p_{XY}p_X p_Y + p_X^2 p_Y^2}{p_{XY}p_X p_Y}\mathrm{d}P_{XY} \\ &= \int\int \left(\frac{p_{XY} - p_X p_Y}{\sqrt{p_{XY}p_X p_Y}}\right)^2 \mathrm{d}P_{XY} \\ &= \mathbb{E}_{p_{XY}}\left[\left(\frac{p_{XY}(x,y) - p_X(x)p_Y(y)}{\sqrt{p_{XY}(x,y)p_X(x)p_Y(y)}}\right)^2\right], \end{aligned} \tag{53}$$

as it is written in (2).

Appendix B: Proof of Theorem 1

The following properties are used for the proof: $\hat{\tilde{\mathbf{p}}}^T\mathbf{1}_N = \hat{\tilde{\mathbf{q}}}^T\mathbf{1}_M = 1$, $\hat{\tilde{\mathbf{J}}}\mathbf{1}_M = \hat{\tilde{\mathbf{p}}}$, $\mathbf{1}_N^T\hat{\tilde{\mathbf{J}}} = \hat{\tilde{\mathbf{q}}}^T$, $[\hat{\tilde{\mathbf{p}}}]\mathbf{1}_N = \hat{\tilde{\mathbf{p}}}$ and $[\hat{\tilde{\mathbf{q}}}]\mathbf{1}_M = \hat{\tilde{\mathbf{q}}}$. Since $\mathbf{1}_N$ and $\mathbf{1}_M$ are the left and right singular vectors of matrix $(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T)$ associated to its null singular value, then we have

$$\left(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T\right)\mathbf{1}_M = \mathbf{0}_N, \quad \mathbf{1}_N^T\left(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T\right) = \mathbf{0}_M^T. \tag{54}$$

From (19) we can write

$$\begin{aligned} \left\|\hat{\tilde{\mathbf{C}}}\right\|^2 = \mathrm{tr}\Bigg(&\left(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T\right)\left(\left[\hat{\tilde{\mathbf{q}}}\right]^{-1} + \frac{\beta}{1-\beta}\mathbf{1}_M\mathbf{1}_M^T\right) \\ &\left(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T\right)^T\left(\left[\hat{\tilde{\mathbf{p}}}\right]^{-1} + \frac{\beta}{1-\beta}\mathbf{1}_N\mathbf{1}_N^T\right)\Bigg) \end{aligned} \tag{55}$$

for any β. From the Woodbury identity, we have

$$\begin{aligned} \left(\left[\hat{\tilde{\mathbf{q}}}\right]^{-1} + \frac{\beta}{1-\beta}\mathbf{1}_M\mathbf{1}_M^T\right)^{-1} &= \left[\hat{\tilde{\mathbf{q}}}\right] - \frac{\left[\hat{\tilde{\mathbf{q}}}\right]\mathbf{1}_M\mathbf{1}_M^T\left[\hat{\tilde{\mathbf{q}}}\right]}{\beta + \mathbf{1}_M^T\left[\hat{\tilde{\mathbf{q}}}\right]\mathbf{1}_M} \\ &= \left[\hat{\tilde{\mathbf{q}}}\right] - \beta\hat{\tilde{\mathbf{q}}}\hat{\tilde{\mathbf{q}}}^T, \end{aligned} \tag{56}$$

$$\begin{aligned} \left(\left[\hat{\tilde{\mathbf{p}}}\right]^{-1} + \frac{\beta}{1-\beta}\mathbf{1}_N\mathbf{1}_N^T\right)^{-1} &= \left[\hat{\tilde{\mathbf{p}}}\right] - \frac{\left[\hat{\tilde{\mathbf{p}}}\right]\mathbf{1}_N\mathbf{1}_N^T\left[\hat{\tilde{\mathbf{p}}}\right]}{\beta + \mathbf{1}_N^T\left[\hat{\tilde{\mathbf{p}}}\right]\mathbf{1}_N} \\ &= \left[\hat{\tilde{\mathbf{p}}}\right] - \beta\hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{p}}}^T. \end{aligned} \tag{57}$$

In the asymptotic case we then have

$$\begin{aligned} \lim_{\beta\to 1}\left(\left[\hat{\tilde{\mathbf{q}}}\right] - \beta\hat{\tilde{\mathbf{q}}}\hat{\tilde{\mathbf{q}}}^T\right)\mathbf{1}_M = \mathbf{0}_M, \\ \lim_{\beta\to 1}\left(\left[\hat{\tilde{\mathbf{p}}}\right] - \beta\hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{p}}}^T\right)\mathbf{1}_N = \mathbf{0}_N. \end{aligned} \tag{58}$$

As a result, these two matrices, which are sample covariance matrices for $\beta \to 1$, share with matrix $(\hat{\tilde{\mathbf{J}}} - \hat{\tilde{\mathbf{p}}}\hat{\tilde{\mathbf{q}}}^T)$ (see (54)) the same singular vectors associated to null singular value. Therefore, the equality with the SMI can be achieved

by computing covariance matrices instead of autocorrelation matrices and using the full-rank matrices $\mathbf{F} \in \mathbb{C}^{N' \times N}$ (with $N' = N - 1$) and $\mathbf{G} \in \mathbb{C}^{M' \times M}$ (with $M' = M - 1$) for the limiting case of $\beta = 1$. For $N' < N - 1$ and/or $M' < M - 1$, however, the smallest singular values will be lost, proving the inequality.

Appendix C: Probability distribution functions of the GMM model

Following (50), the joint distribution turns out to be a GMM with

$$\begin{aligned} p_{XY}(x,y) =& \frac{1}{4}\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \frac{1}{4}\mathcal{N}(-\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ &+ \frac{1}{4}\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + \frac{1}{4}\mathcal{N}(-\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \end{aligned} \tag{59}$$

where

$$\begin{aligned} \boldsymbol{\mu}_1 = \begin{bmatrix} -\sqrt{\lambda\rho} & \sqrt{\lambda\rho} \end{bmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} \sqrt{\lambda\rho} & \sqrt{\lambda\rho} \end{bmatrix}, \\ \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1-\lambda\rho & (1-\lambda)\rho \\ (1-\lambda)\rho & 1-\lambda\rho \end{bmatrix}, \\ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1-\lambda\rho & -(1-\lambda)\rho \\ -(1-\lambda)\rho & 1-\lambda\rho \end{bmatrix}. \end{aligned} \tag{60}$$

The marginal probability density functions are the following:

$$\begin{aligned} p_X(x) &= \frac{1}{2}\mathcal{N}\left(\sqrt{\lambda\rho}, 1-\lambda\rho\right) + \frac{1}{2}\mathcal{N}\left(-\sqrt{\lambda\rho}, 1-\lambda\rho\right), \\ p_Y(y) &= \frac{1}{2}\mathcal{N}\left(\sqrt{\lambda\rho}, 1-\lambda\rho\right) + \frac{1}{2}\mathcal{N}\left(-\sqrt{\lambda\rho}, 1-\lambda\rho\right). \end{aligned} \tag{61}$$

Note that ρ is equivalent to a correlation coefficient, but constrained within $0 \leq \rho < 1$, where 1 would be equivalent to infinite SMI, and 0 to null SMI. On the contrary, λ defines the implicit probability distributions and can be used for generating different degrees of difficulty in terms of SMI estimation with $0 \leq \lambda \leq 1$.

Appendix D: Perturbation variance setting

For large L, the bias and variance of the SMI estimator are given by

$$\text{bias}\left(\hat{I}_s\right) = -O\left(\sigma^2\right) + O\left(\sigma^{-1}L^{-1}\right) \tag{62}$$

$$\text{var}\left(\hat{I}_s\right) = O\left(\sigma^{-1}L^{-1}\right). \tag{63}$$

The term $O(\sigma^2) \geq 0$ is a result of the data processing inequality and the consistent (with L) terms $O(\sigma^{-1}L^{-1})$ decrease with σ as a result of (34). Both have also been approximately confirmed by simulations for a wide range of scenarios. As the mean squared error is $\text{mse}(\hat{I}_s) = \text{bias}^2(\hat{I}_s) + \text{var}(\hat{I}_s)$, the condition $\lim_{L\to\infty} \sigma^2 = \lim_{L\to\infty} \sigma^{-1}L^{-1} = 0$ is required to yield $\lim_{L\to\infty} \text{mse}(\hat{I}_s) = 0$, which moves to choosing σ as a monotonically decreasing function of L such that $\sigma^{-1}L^{-1}$ is also monotonically decreasing. Let us adopt a power law $\sigma = O(L^{-\gamma})$, for which the condition $0 < \gamma < 1$ guarantees the desired convergence given by

$$\text{mse}\left(\hat{I}_s\right) = O\left(L^{-\min[4\gamma, 1-\gamma]}\right). \tag{64}$$

Then, the value of γ can finally be optimized by the following MiniMax rule:

$$\gamma = \arg\max_{\gamma} \left(\min\left[4\gamma, 1-\gamma\right]\right) = \frac{1}{5}. \tag{65}$$

The resulting power law acts similar to the Silverman's rule for kernel smoothing, which implies setting the perturbation variance as $\sigma^2 = p/\left(L^{2/5}\right)$, being p the new relative free parameter of the estimator.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: Wiley, 2006.

[2] Q. Wang, S. R. Kulkarni, and S. Verdú, *Universal estimation of information measures for analog sources*. Foundations and trends in Communications and Information Theory, 2009, no. 5:3.

[3] S. Verdú, "Empirical estimation of information measures: A literature guide," *Entropy, 21(8), 720*, pp. 1–16, July 2019.

[4] J. C. Príncipe, *Information Theoretic Learning: Rényi's Entropy and Kernel Perspectives*. NewYork: Springer, 2010.

[5] K.-C. Chen, S.-L. Huang, L. Zheng, and H. V. Poor, "Communication theoretic data analytics," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 4, pp. 663–675, Apr. 2015.

[6] Y. C. Eldar, A. O. Hero III, L. Deng, J. Fessler, J. Kovacevic, H. V. Poor, and S. Young, "Challenges and open problems in signal processing: Panel discussion summary from ICASSP 2017," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 8–23, Nov. 2017.

[7] Y.-I. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Physical Review E*, vol. 52, no. 3, pp. 2318–2321, 1995.

[8] P. Hall and S. C. Morton, "On the estimation of entropy," *Annals of the Institute of Statistical Mathematics*, no. 15, pp. 1419–1519, 1993.

[9] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[10] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E.*, vol. 69, no. 6, 2004.

[11] A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. Wasserman, "Nonparametric estimation of Rényi divergence and friends," *International Conference on Machine Learning*, PMLR, pp. 919–927, 2014.

[12] M. Sugiyama, "Machine learning with squared-loss mutual information," *Entropy*, vol. 15, no. 1, pp. 80–112, 2013.

[13] F. de Cabrera and J. Riba, "Squared-loss mutual information via high-dimension coherence matrix estimation," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 5142–5146, 2019.

[14] S.-L. Huang, C. Suh, and L. Zheng, "Euclidean information theory of networks," *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6795–6814, 2015.

[15] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "On universal features for high-dimensional learning and inference," arXiv preprint arXiv:1911.09105 (2019).

[16] X. Xu, S.-L. Huang, L. Zheng and G. W. Wornell. "An Information Theoretic Interpretation to Deep Neural Networks," *Entropy 2022*, 24, 135.

[17] F.d.P. Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen and K.R. Duffy, "Principal Inertia Components and Applications," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 5011–5038, 2017.

[18] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Mach. Learning Res.*, vol. 3, pp. 1–48, 2002.

[19] M. Espinoza, J.A.K. Suykens and B. De Moor, "Least squares support vector machines and primal space estimation," *42nd IEEE International Conference on Decision and Control*, vol. 4, pp. 3451–3456, 2003.

[20] M. Braun, J. Buhmann, and K. Müller, "On relevant dimensions in kernel feature spaces," *Journal of Machine Learning Research*, vol. 9, pp. 1875 – 1908, 2008.

[21] D. Ramírez, J. Vía, I. Santamaría, and L. L. Scharf, "Locally most powerful invariant tests for correlation and sphericity of Gaussian vectors," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2128–2141, 2013.

[22] Z. Goldfeld, K. Greenewald, J. Weed, and Y. Polyanskiy, "Optimality of the plug-in estimator for differential entropy estimation under gaussian convolutions," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 892–896, 2019.

[23] D. Ramírez, J. Vía, I. Santamaría, and P. Crespo, "Entropy and Kullback-Leibler divergence estimation based on Szegö's theorem," in *European Signal Processing Conference, EUSIPCO*, 2009.

[24] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler digergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[25] S. Tridenski, R. Zamir, and A. Ingber, "The Ziv-Zakai-Rényi bound for joint source-channel coding," *IEEE Transactions on Information Theory*, vol. 61, no. 8, pp. 4293–4315, 2015.

[26] S. Verdú, "α-mutual information," in *Information Theory and Applications Workshop (ITA)*, 2015.

[27] M. Tomamichel and M. Hayashi, "Operational interpretation of Rényi information measures via composite hypothesis testing against product and Markov distributions," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1064–1082, 2018.

[28] A. Lapidoth and C. Pfister, "Two measures of dependence," *Entropy*, vol. 21, 2019.

[29] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese, "Mutual information estimation reveals global associations between stimuli and biological processes," *BMC Bioinformatics*, vol. 10, no. 1, pp. 1–12, 2009.

[30] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, no. 4, pp. 417–528, 2004.

[31] K. Pearson, *On the theory of contingency and its relation to association and normal correlation*. Dulau and Company, 1904.

[32] A. Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[33] X. S. K. Fukumizu, A. Gretton and B. Schölkopf, "Kernel measures of conditional dependence," *Advances in Neural Information Processing Systems*, 2008.

[34] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, 2006.

[35] G. Govaert and M. Nadif, "Mutual information, phi-squared and model-based co-clustering for contingency tables," *Advances in Data Analysis and Classification*, vol. 12, no. 3, pp. 455–488, 2018.

[36] S. Kotz and S. Nadarajah, *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

[37] G. A. Darbellay and I. Vajda, "Entropy expressions for multivariate continuous distributions," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 709–712, 2000.

[38] E. Lukacs, *Characteristic functions*. Griffin, 1970.

[39] L. L. Scharf and C. T. Mullis, "Canonical coordinates and the geometry of inference, rate, and capacity," *IEEE Trans. Signal Process.*, vol. 48, no. 3, pp. 824–831, 2000.

[40] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," arXiv preprint arXiv:1304.6133 (2013).

[41] S.-L. Huang and L. Zheng, "Linear information coupling problems," in *IEEE Int. Symp. on Information Theory*, pp. 1029–1033, 2012.

[42] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf, "Kernel methods for measuring independence," *Journal of Machine Learning Research*, 2005.

[43] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 1936.

[44] A. Pezeshki, L. L. Scharf, M. R. Azimi-Sadjadi, and M. Lundberg, "Empirical canonical correlation analysis in subspaces," in *38th Asilomar Conf. Signals, Syst. Comput.*, vol. 1, pp. 994–997, 2004.

[45] S. Haykin, *Neural networks and learning machines*, 3rd ed. Prentice-Hall, 2009.

[46] J.-L. Rojo-Álvarez, M. Martínez-Ramón, J. Muñoz-Marí, and G. Camps-Valls, *Digital Signal Processing with Kernel Methods*, Wiley-Blackwell, Ed. IEEE Press, 2018.

[47] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Technical Report CSD-TR-03-02, Royal Holloway University of London*, 2003.

[48] S. Csörgő, "Testing for independence by the empirical characteristic function," *Journal of Multivariate Analysis*, vol. 16, no. 3, pp. 290–299, 1985.

[49] G. J. Székely, M. L. Rizzo, and N. K.Bakirov, "Measuring and testing dependence by correlation of distances," *The annals of statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[50] F. de Cabrera and J. Riba, "A novel formulation of independence detection based on the sample characteristic function," in *26th European Signal Processing Conference (EUSIPCO)*, pp. 2608–2612, 2018.

[51] S. M. Kay, *Modern Spectral Estimation: Theory and Application*. Prentice Hall, 1988.

[52] D. Brillinger, "The key role of tapering in spectrum estimation," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 5, pp. 1075–1076, 1981.

[53] J.-F. Collet, "An exact expression for the gap in the data processing inequality for f-divergences," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4387–4391, 2019.

[54] A. Feuerverge and R. A. Mureika, "The empirical characteristic function and its applications," *The annals of Statistics*, vol. 5, no. 1, pp. 88–97, 1977.

[55] A. Papoulis and H. Saunders, *Probability, random variables and stochastic processes*. McGraw-Hill, NewYork, 1989.

[56] U. Grenander and G. Szegö, *Toeplitz forms and their applications*. Berkeley: Univ. Calif. Press, 1958.

[57] R. Gray, *Toeplitz and circulant matrices: a review*. Foundations and trends in Communications and Information Theory, 2006.

[58] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[59] B. Silverman, *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.

[60] S. Chen, "Optimal bandwidth selection for kernel density functionals estimation," *Journal of probability and statistics*, vol. 2015, no. ID 242683, pp. 1–22, 2015.

[61] C. L. Haley and M. Anitescu, "Optimal bandwidth for multitaper spectrum estimation," *IEEE Signal Processing Letters*, vol. 45, no. 11, pp. 1315–1321, 2017.

[62] M. Noshad, J. Choi, Y. Sun, A. Hero and I.D. Dinov, "A data value metric for quantifying information content and utility," *Journal of Big Data*, vol. 8, 2021.

[63] J. F. Silva and S. Narayanan, "Complexity-regularized tree-structured partition for mutual information estimation," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1940–1952, 2012.

[64] R. Malladi, D. H. Johnson, G. P. Kalamangalam, N. Tandon, and B. Aazhang, "Mutual information in frequency and its application to measure cross-frequency coupling in epilepsy," in *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 3008–3023, 2018.

[65] C. A. López, F. de Cabrera, and J. Riba, "Estimation of information in parallel Gaussian channels via model order selection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5675–5679, 2020.

Ferran de Cabrera received a B.E. degree in telecommunication sciences and technologies from the Universitat Politècnica de Catalunya (UPC), Spain, in 2015, and a M.Sc. degree in telecommunications engineering from the Universitat Politècnica de Catalunya (UPC), Spain, in 2017. In 2018, he joined the Signal Theory and Communications Department, Universitat Politècnica de Catalunya (UPC), where he is currently working toward a Ph.D. degree in signal theory and communications. His current research interests include statistical signal processing and signal processing algorithms for communications and machine learning.

Jaume Riba (SM'05) received the M.Sc. and Ph.D. degrees in telecommunications engineering from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1992 and 1997, respectively. In 1992 he entered the Department of Signal Theory and Communications (UPC) as Assistant Professor and was promoted to Associate Professor in 1997. His current research interests are in the area of signal processing for communications, array processing, synchronization, localization techniques, and measures of entropy and information. He served as Associate Editor of the IEEE Tr. Sig. Pro. and as Guest Editor in a special issue of IEEE Sig. Pro. Magazine. He has been involved in several signal processing research and development projects in the framework of the ESA programs. He is a co-recipient of the 2003 IEEE Best Paper Award from the IEEE Signal Processing Society, for the paper entitled "Conditional Maximum Likelihood Timing Recovery: Estimators and Bounds", and a co-recipient of the IEEE Best Paper Award for the Signal Processing for Communications Symposium of IEEE, (ICC, 2013), for the paper entitled "Sampling walls in signal detection of Bernoulli nonuniformly sampled signals".